Data-driven, memory-based computational models of human segmentation of musical melody

Miguel Ferrand **Amoroso Lopes**

Doctor of Philosophy
Music
School of Arts, Culture and Environment
University of Edinburgh
2005

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Miguel Ferrand*)

Abstract

When listening to a piece of music, listeners often identify distinct sections or segments within the piece. Music segmentation is recognised as an important process in the abstraction of musical contents and researchers have attempted to explain how listeners perceive and identify the boundaries of these segments.

The present study seeks the development of a system that is capable of performing melodic segmentation in an unsupervised way, by learning from non-annotated musical data. Probabilistic learning methods have been widely used to acquire regularities in large sets of data, with many successful applications in language and speech processing. Some of these applications have found their counterparts in music research and have been used for music prediction and generation, music retrieval or music analysis, but seldom to model perceptual and cognitive aspects of music listening.

We present some preliminary experiments on melodic segmentation, which highlight the importance of memory and the role of learning in music listening. These experiments have motivated the development of a computational model for melodic segmentation based on a probabilistic learning paradigm.

The model uses a Mixed-memory Markov Model to estimate sequence probabilities from pitch and time-based parametric descriptions of melodic data. We follow the assumption that listeners' perception of feature salience in melodies is strongly related to expectation. Moreover, we conjecture that outstanding entropy variations of certain melodic features coincide with segmentation boundaries as indicated by listeners.

Model segmentation predictions are compared with results of a listening study on melodic segmentation carried out with real listeners. Overall results show that changes in prediction entropy along the pieces exhibit significant correspondence with the listeners' segmentation boundaries.

Although the model relies only on information theoretic principles to make predictions on the location of segmentation boundaries, it was found that most predicted segments can be matched with boundaries of groupings usually attributed to Gestalt rules.

These results question previous research supporting a separation between learning-based and innate bottom-up processes of melodic grouping, and suggesting that some of these latter processes can emerge from acquired regularities in melodic data.

Acknowledgments

I wish to express my gratitude to:

My supervisors Peter Nelson and Geraint Wiggins for many insightful conversations during the development of this research, and for their advice, guidance and encouragement to bring this project to a conclusion.

Alan Smail, Darrell Conklin, Marcus Pierce, David Meredith, Taylan Cemgil, Richard Shillcock, Petri Toivianen, Emilios Cambouropoulos, Mark Steedman, Raymond Monelle, for lending some of their time and expertise.

All participants in the listening experiments for their contribution to this research

The staff in Music, Rob Dow, Katrina Joyce, Natalie Caron, Scott Walker for being always so helpful

ISMS Group at Goldsmith University of London (previously at City University) for their logistic and scientific support

My examiners, Henkjan Honing and Katie Overy, for their useful comments and suggestions on this thesis.

I am especially indebted to:
Christina Anagnostopoulou, Dimitra Tripany, Running Bear, Rosalía Vázquez, Jordan Fleming, Martin Parker, Anna Ritta Adessi, João Leite, Francisco Pereira and so many others for their friendship and support

My mother Clara, brother João and sister Maria for their love and great support

Leigh for sharing this journey with me in good and difficult times, with great patience and love.

This research was supported by a 3-year EPSRC grant.

Contents

1 Introduction 1

1.1 Motivation . . . 1
1.2 Aims and Contributions . . . 3
1.3 Outline of the Dissertation . . . 4

2 Background 7

2.1 Perceptual and Cognitive Factors in Melodic Segmentation . . . 7
2.2 Learning and Expectancy in Music Perception . . . 16
2.3 Existing Models of Music Segmentation . . . 21
2.4 Summary . . . 27

3 Information Theory and Probabilistic Modelling 29

3.1 Information Theory . . . 29
3.2 Probabilistic (Language) Modelling . . . 33

4 Preliminary Experiments on Melodic Segmentation 39

4.1 Introduction . . . 39
4.2 Melodic Density Segmentation Model . . . 41
4.3 Experiments and Results . . . 44
4.4 Discussion . . . 46
4.5 Summary . . . 48

5 An Empirical Study of Melodic Segmentation 55

5.1 Preliminary Segmentation Studies with Listeners . . . 55
5.2 Description of the Study . . . 57
5.3 Results . . . 60
5.4 Discussion . . . 67
5.5 Summary . . . 69

6 A Probabilistic Model of Melody Segmentation **71**
6.1 Melodic Representation 71
6.2 The Memory Model 73
6.3 Segment Boundary Prediction 77
6.4 Summary 79

7 Experimental Results **81**
7.1 Methodology 81
7.2 Boundary Prediction Results 84
7.3 Discussion of results 99
7.4 Summary 101

8 Discussion and Related Work **103**
8.1 Methodological Considerations 103
8.2 The Influence of the Representation 104
8.3 A Probabilistic Memory Model 106
8.4 Modelling similarity 108
8.5 Feature Learning and Gestalt 109
8.6 Summary 111

9 Conclusion **113**
9.1 Summary of Contributions 113
9.2 Future Work 115

A Event lists of melodies used in the listening study **119**

B Kernel Density Estimation **127**

C Listening study **129**

D Experimental Results Data **139**

E Supplementary material **151**

Bibliography **153**

List of Figures

2.1 Illustration of some Gestalt principles 8

3.1 Comparing *information* and *souce entropy* for a process based on a fair coin toss and a biased coin toss 31
3.2 Entropy for a binary memory-less source, where two possible outcomes have probabilities p and $1-p$ 31
3.3 Comparison between n-gram and Mixed-memory models on the acquisition of symbol dependencies. 36

4.1 LBDM and MDSM boundary profiles for melody *Frère Jacques* 51
4.2 LBDM and MDSM boundary profiles for the theme of Mozart's Symphony in Gm 52
4.3 53

5.1 Screen-shot of the Music Puncher interface during a session 59
5.2 Example of data logged by one participant in a listening session 61
5.3 Segment probability density estimation for melody f0927 64
5.4 Segment probability density and score of melody f0927 65
5.5 Segment probability density and score of melody K284 66
5.6 Segment probability density and score of melody E0547 68

6.1 Example of melodic representation for the first two bars of Syrinx 74
6.2 Event list for the opening theme of Mozart's Symphony in Gm 75
6.3 Transition tables and mixture coefficients for PS and DR (1st and 2nd order) for the opening theme of Mozart's Symphony in Gm 76
6.4 Example of boundary predictions for the opening theme of Mozart's Symphony in Gm 80

7.1 Comparison between model boundary predictions and listeners' boundaries for Syrinx . . . 87
7.2 Examples of event sequences representing ornaments in Syrinx . . . 89
7.3 Boundary salience profiles for melody K284 . . . 91
7.4 Comparison between model's boundary predictions and listeners' boundaries for melody k284 . . . 92
7.5 Entropy profiles $H_c(PS)$ and $H_c(DR)$ for folk-song E0547 . . . 95
7.6 Boundary salience profiles for folk-song E0547 . . . 97

A.1 Event list for Syrinx . . . 120
A.2 Event list for K284 (extract) . . . 121
A.3 Event list for K333 (extract) . . . 122
A.4 Event list for folk-song E0547 . . . 123
A.5 Event list for folk-song F0927 . . . 124
A.6 Event list for folk-song Q0034 . . . 125

C.1 Overview of the participants in the listening study . . . 130
C.2 Copy of instruction sheet/questionnaire handed to the participants. . . . 131
C.3 Probability density of listeners' segment boundaries for Syrinx . . . 132
C.4 Probability density of listeners' segment boundaries for K284 . . . 133
C.5 Probability density of listeners' segment boundaries for K333 . . . 134
C.6 Probability density of listeners' segment boundaries for E0547 . . . 135
C.7 Probability density of listeners' segment boundaries for F0927 . . . 136
C.8 Probability density of listeners' segment boundaries for Q0034 . . . 137

D.1 Boundary salience profiles for Syrinx . . . 141
D.2 Boundary salience profiles for melody K333 . . . 143
D.3 Comparison between model boundary predictions and listeners' boundaries for melody K333 . . . 144
D.4 Comparison between model boundary predictions and listeners' boundaries for melody E0547 . . . 145
D.5 F0927: boundary predictions $S(C)$ and successor probability $P(X|C)$ for features PS and DC. Boundary selection threshold indicated by a dotted line at the bottom of the graph . . . 146
D.6 Comparison between model boundary predictions and listeners' boundaries for melody F0927 . . . 147

D.7 Comparison between model boundary predictions and listeners' boundaries for melody Q0034 148
D.8 Folk-song q0034: average entropy $H(c)$ and outcome probability $P(x|c)$. 149
D.9 Boundary salience profiles for folk-song Q0034 150

List of Tables

4.1 *Order* and *recency* of pitch intervals for a sequence of events. Intervals are in semitones. 42
4.2 Interval frequencies and ratings for a major/minor scale framework . . . 43
4.3 Segmentation results obtained for 7 melodies, comparing the LBDM and the MDSM 45
4.4 F-measure for the LBDM and MDSM 46

5.1 Participants' age and gender mix 57
5.2 Description of the melodies used in the listening study 58
5.3 Statistical analysis of boundary segment counts for musician and non-musician subjects 61
5.4 Correlations between segment boundary probability density profiles for musician and non-musician subjects 64

7.1 Characterisation of boundary predictions for Syrinx 85
7.2 Listeners' boundaries for Syrinx 85
7.3 Duration ratio representation for four occurrences of Syrinx' opening motif 88
7.4 Characterisation of boundary predictions for melody K284 90
7.5 Characterisation of boundary predictions for melody K333 93
7.6 Characterisation of boundary predictions for melody E0547 94
7.7 Characterisation of boundary predictions for melody f0927 96
7.8 Characterisation of boundary predictions for melody Q0034 98
7.9 Summary of boundary prediction results, for the two threshold levels considered. 100
7.10 Characterisation of correct boundary predictions according to source feature and underlying melodic principle 101

B.1 Kernel functions 128

D.1 Selected listener boundaries for Syrinx 140
D.2 Selected listener boundaries for K284 140
D.3 Selected listener boundaries for K333 142
D.4 Selected listener boundaries for melody E0547 142
D.5 Selected listener boundaries for melody F0927 145
D.6 Selected listener boundaries for melody Q0034 147

Chapter 1

Introduction

"In general music is always hard to understand unless it is made easier by repetition of as many minute, small, medium or large sections as possible. The first precondition for understanding is, after all memory..." (Schoenberg, 1975)

1.1 Motivation

Researchers have been trying to understand how humans perceive music and how they build mental schemas of the musical works to which they listen. When listening to a piece of music, listeners often identify distinct sections or segments within the piece, therefore music segmentation is likely to be an important underlying process in the abstraction of musical contents. This is conveyed in several music theories (Lerdahl and Jackendoff, 1983; Narmour, 1992; Cambouropoulos, 1998) which are based on the general assumption that an important part of music understanding relies on the segmentation of a piece into constituent units.

Segmentation in the visual or auditory domain, seems to be a fundamental part of human processing of sensorial data and is often seen as a form of perceptual economy. Research on music segmentation (Lerdahl and Jackendoff, 1983; Bregman, 1990; Cambouropoulos, 1998; Temperley, 2001) makes frequent references to the theories of Gestalt psychology (Wertheimer, 1938), which proposed a series of laws to explain how the mind associates perceived stimuli by forming groups or patterns. These may be applied in the music domain to identify discontinuities and create groupings between musical events. It has often been conveyed that Gestalt principles operate independently of the listeners musical knowledge but it has been suggested (Meyer, 1956) that learning might condition the operation of these laws making them more difficult to generalise.

The perception of parallelism and similarities are also known to influence the listener's ability to identify different sections within a musical piece (Deliège, 1997). When familiarised with a given musical piece, listeners memorise recurrent features in the music and are likely to use this knowledge to carry out further musical analytical tasks. Empirical evidence has shown that large sections of a musical piece can be recalled by listeners based on the recurrence of smaller musical cells (Deliège and Melén, 1997), which act as salient markers within the piece. These findings are important because they suggest that despite memory limitations, melodic similarity relations can be established from a reduced (or prototypical) set of recollections from a piece.

In his theory on music expectation Meyer (1956) outlines the importance of learning in music understanding and relates expectation with information theoretical notions such as entropy (Meyer, 1967). In a sequence of events, the predictability (or unpredictability) associated with the occurrence of a musical event can change its prominence and make it salient to the listener. This suggests that the salience of a musical event is likely to be context-dependent and may be associated with musical features that present intra-opus or inter-opus distinctiveness (Huron, 2001).

The cognitive relevance of the frequency distribution of musical elements has been extensively discussed by Krumhansl who reports that "listeners appear to be very sensitive to the frequency with which the various elements and their successive combinations are employed in music" (1990, p. 286). This suggests that key structural elements, possibly underlying segment boundaries, may be learned from the statistical properties of musical data.

Probabilistic models have been widely used to capture the regularities in large sets of data, with many successful applications in natural language and speech processing (Manning and Schüttze, 1999). Some of these methods have migrated into the music domain however, probabilistic learning models have been used mostly for music generation (Cope, 1991; Conklin and Witten, 1995; Ponsford et al., 1999; Reis, 1999) and seldom to model musical analytical tasks such as segmentation (Bod, 2001).

Melodic segmentation can be a useful and complementary processing technique for information retrieval. Pattern induction techniques, for example, can benefit from "methods that are geared towards finding perceptually-pertinent local boundaries" (Cambouropoulos et al., 2001). These local boundaries can be used to filter the results of a pattern search process providing a frame of reference for the selection of 'musically significant' segments (Meredith et al., 2002).

The emphasis on the perceptual pertinence of a segmentation is a central motivation

behind the models developed in the course of this work. This quality of the boundaries, is paramount for complex information retrieval problems such as query-by-humming (Birmingham et al., 2001) where human input is error-prone. In these cases some initial abstraction of a piece is needed, as often pieces of music are too long to be searched efficiently.

Computer models are valuable exploratory tools which allow us to look at large quantities of data in a systematic way. The study of music perception by means of computational modelling is attractive because it forces us to analyse and make explicit certain cognitive processes, as we attempt to model these. So in order to develop a model for melodic segmentation it is paramount to obtain reference segmentation data obtained from human subjects. Without reliable data about human behaviour the psychological value of computational models is always limited (Desain et al., 1998).

1.2 Aims and Contributions

The main goal of this research is to develop a computational model capable of generating perceptually pertinent segmentations for a given melody, with minimum prior knowledge given to the system other than the melody input. To achieve this we propose to address the following:

- Memory-based melodic segmentation. A review of existing literature on music perception and cognition will highlight the importance of memory and learning in experiencing music. We propose to develop two computational models for automatic melodic segmentation, which model some of the effects of memory in music listening.

  The first approach is a short-term memory segmentation model that aims to simulate the effects of recency to rate the salience of segment boundaries. The model embodies a notion of melodic density over time and follows the conjecture that boundaries are likely to be perceived where melodic density is low. The performance of this model is compared with an existing segmentation model, the LBDM (Cambouropoulos, 1998), for a set of melodic examples and it will be shown that an improvement in boundary selectivity can be achieved.

  The second approach uses a probabilistic learning paradigm to acquire regularities from a given melody and then uses the stored probabilistic information to make predictions about the locations of segment boundaries. We conjecture that salient

melodic features are retained due to repeated listening of a musical repertoire (or even of a single piece) and that these provide important cues to segmentation. Segment prediction is based on the assumption that perceived salient boundaries in a melody are associated with changes in the predictability of certain melodic features.

- Learning from non-annotated musical data. This is a central aspect of this research meaning that the input to our models is an event-based parametric representation of a melody (i.e. MIDI). By working with raw melodic data, we exclude information obtained from a score or any other annotations introduced by an expert. This, as will be shown, is in contrast with other existing models of music perception. In this research we adopt a bottom-up approach to music perception where melodic constructs or abstract rules are to be implicitly derived from the melodic data, thus pursuing a more realistic model of a listening experience.

- Evaluate results with real listeners. One of the aims of this research is to compare the results of our segmentation model with segmentation data obtained in a real listening setting. Due to the lack of available segmentation data, a listening study was devised in order to produce segmentation data for a set of test melodies. Participants in this study included both trained musicians and subjects with no specific musical training. This listening study provided valuable data for the evaluation of our computational segmentation model but also provided some insight on how previous musical knowledge may influence a listener's ability to perform simple musical analytical tasks such as segmentation. From the comparison of the listeners and the model's segmentation we will discuss how a machine learning paradigm may embody implicitly some aspects of melodic perception such as motivic similarity and Gestalt-based note groupings, often defined explicitly in other segmentation models.

1.3 Outline of the Dissertation

This dissertation is organised as follows:

Chapter 2 presents a review of research on music perception and cognition, associated with the problem of melodic segmentation. This chapter also includes a critical overview of existing models of melodic segmentation.

Chapter 3 complements Chapter 2 and presents a more specialised overview on Information Theory and Probabilistic modelling which constitutes the theoretical background for Chapter 6.

Chapter 4 describes preliminary work towards a memory-based model for melodic segmentation. A segmentation model is developed based on the notion of melodic density, emphasising the role of short-term memory and time in music listening, by modelling the effects of recency in the perception of boundaries. This chapter reports some experimental results obtained by comparing this model with the Local Boundary Detection Model (Cambouropoulos, 1998, 2001a) for a set of melody examples.

Chapter 5 presents an empirical study of melodic segmentation where listeners, both with and without formal music training, were asked to segment several melodies while listening to them.

Chapter 6 describes in detail a probabilistic model of melodic segmentation, addressing the adopted melodic representation and memory model, and presenting the hypothese that underlie our segment boundary prediction method.

Chapter 7 presents the experimental results obtained with the segmentation model described in Chapter 6 by comparing them with the melodic segmentation data obtained with listeners (presented in Chapter 5).

Chapter 8 presents a general discussion of the results and a critical analysis of our probabilistic model of melodic segmentation, addressing its strengths and weaknesses, and relating it to existing models and other theories on music perception and cognition.

Chapter 9 concludes this dissertation by summarising the main contributions of this research work before identifying themes for future work.

Some of the research material presented in this dissertation has been partially included in some publications (Ferrand et al., 2002, 2003a,b).

Chapter 2

Background

Segmentation is inherent to most human information processing and is often seen as a form of perceptual economy. In this chapter we review theoretical and empirical research work that relates to melodic segmentation.

First we identify and describe perceptual and cognitive factors which are believed to affect listeners in the abstraction of music and consequently in melodic segmentation. We discuss the role of Gestalt principles of perception and discuss their relevance in music perception. We also address the role of musical parallelism in segmentation and look at factors that may enhance or inhibit the perception of similarities in melodies.

Then we discuss the role of memory in music perception, particularly the influence of learning and the role of expectancy in music listening, referring to existing theories, and showing how these factors may influence the perception of segment boundaries in music.

In the final section we review existing computational models of melodic segmentation.

2.1 Perceptual and Cognitive Factors in Melodic Segmentation

2.1.1 Gestalt Psychology in Music

The theories of Gestalt psychology, attributed to Wertheimer (1938), proposed a series of laws, aimed to explain how the mind associates these perceived stimuli, by forming groups or patterns. The word 'gestalt' is often translated as 'shape', 'unified whole' or 'pattern' and refers to the way we derive meaningful components from perceived stimuli.

A few conditions are necessary for the formation of such groupings or 'gestalts'. Some of these conditions can be summarised as follows:

Similarity Elements that appear to be similar will tend to be grouped together;

Proximity Elements that are close together are likely to be perceived as a group;

Closure Elements tend to form a group if they appear to complete a pattern;

Good Continuation Elements that follow an established pattern are more likely to be grouped with elements that continue the pattern than with those that deviate from it;

Common fate Elements that move together in similar direction tend to be perceived as a group rather than as individual elements;

Most descriptions and examples found in the literature to illustrate the Gestalt laws are commonly of a graphical nature (as in Figure 2.1), even when referring to non-visual phenomena. But although the Gestalt principles were initially presented in the context of visual perception, Wertheimer (1938) did suggested that these laws could be applied in the auditory domain, presenting a few examples where chromatic sequences of tones were used to illustrate the principles of proximity and continuity.

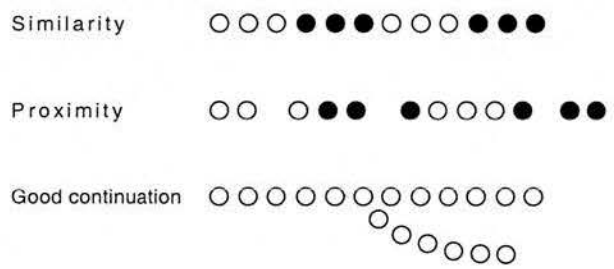


Figure 2.1: Illustration of some Gestalt principles

One of the problems with Gestalt laws is that it is not always easy to define the boundaries between some of their underlying principles. This is particularly true when we transfer the principles to the auditory domain. Bregman (1990) observes that it is particularly hard to know where proximity principle leaves off and the similarity principle begins. For example, he writes, "when we say that two frequencies are near each other we are employing a spatial analogy, but we could just as easily have said that they are similar to one another" Bregman (1990, p. 198). Bregman notes that perhaps it is

not relevant to identify which principle applies, as long as the effects of the two are the same. Other types of ambiguity can appear when Gestalt principles are conflicting. For example, a sequence of notes of short duration immediately followed by a sequence of long notes may be segmented in two distinct groups of notes, either based on 'proximity', grouping notes equidistant in time, or based on 'similarity' grouping notes with identical duration. This suggests that boundaries may be the product of several different grouping forces, and so it is expected that groups may overlap or have ambiguous boundaries between them.

Many of the Gestalt laws were considered by their proponents as different forms of the more general Law of Prägnanz which states that, of all possible organisations, the one which will occur has the most stable and simple shape. More recently, it has been argued (Rock and Palmer, 1990) that all Gestalt principles can in fact be different forms of a principle of similarity. So 'proximity' can be seen as similarity in location, 'common fate' as similarity in time, etc. The authors suggest that Gestalt principles can generically take the form:

> "All else being equal, elements that are related by X tend to be grouped perceptually into higher-order units",

where X is a property or some association criterion. This rather compact definition of the Gestalt principles fails to specify what characteristics of the perceptual elements can be judged as similar. Furthermore, if different criteria can be used to relate a given set of elements, how are they to be combined?

The Gestalt principles have been the focus of much debate. At the center of the controversy is their ambiguous nature. While some postulate that Gestalt principles of perception are innate and/or universal, others have argued that learning is intimately related to the way they are used in perception.

While some postulated that innate perceptual primitives are responsible for some Gestalts, some distinction is necessary in the multitude of laws that have been presented in the literature. It seems that different Gestalt laws might rely on different perceptual processes and therefore might be different in nature. For example, defining a rule for 'proximity' seems by far less controversial that defining a rule for 'good form'. This idea is corroborated by Narmour (1992) who argues that empirically, similarity, proximity, and common direction can be measured and defined in a rigourous sense, whereas notions of 'good' and 'best' cannot.

Reybrouck (1997) writes that "at the lowest level there are systems for recognising patterns whose information is contained in the external world. Assigning these patterns to music, however, cannot be done on a priori-grounds, for the delimitation of musical patterns is greatly influenced by the selection processes and decoding mechanisms of the listener."

The validity of some of the principles associated to the so called Gestalt-laws has in fact been corroborated by empirical findings as described by Bregman (1990). As Leman and Schneider (1997) point out, there have been "too many Gestalt laws, and perhaps not enough hardcore explanations to account for these, not withstanding the great amount of experimental work that had been done". Meyer (1956) makes a distinction between the experimental findings made in connection with Gestalt theory and the theory itself. This distinction shows that, as Meyer writes, it is possible to accept the empirical data, the laws, discovered by Gestalt psychologists without adopting the hypothetical explanations furnished by the theory" (Meyer, 1956). Meyer argues for the importance of learning in perception, and criticises the idea that spontaneous natural laws of organisation can account for all grouping phenomena. If we accept that "the mind organises and groups stimuli it perceives, into the simplest possible shapes or the most satisfactory" we realise that what may be considered a simple shape or the most satisfactory organisation is a product of cultural experience. Meyer emphasises, with a few examples, the role of learning in conditioning the operation of the Gestalt laws conveying the idea that it is difficult to generalise Gestalt concepts to explain all aspects of musical perception (Meyer, 1956). He concludes that "although there is ample reason to believe that the laws developed by Gestalt psychologists, largely in connection with visual experience, are applicable in a general way to aural perception, they cannot be made the basis of a thoroughgoing system for the analysis of musical perception and experience".

Gestalt principles such as proximity, similarity and good continuation have been included in several musical theories and cognitive models of music (Lerdahl and Jackendoff, 1983; Narmour, 1992; Cambouropoulos, 1998; Temperley, 2001) and were shown to have reasonable predictive power in musical analytical tasks (Krumhansl, 1990), such as the establishment of groupings and segmentation of musical works.

2.1.2 Melodic Similarity

Similarity is a fundamental phenomenon in human perception and cognition, and it has been associated with cognitive abilities such as problem solving, learning, memorisation

and attention (Goldstone and Son, 2005). The complexity of human similarity judgements has motivated considerable research in the fields of psychology and cognitive science.

Early attempts to define and measure similarity make use of geometrical models (Goldstone, 1999) where entities are represented as points in a multi-dimensional space, usually as a result of pair-wise comparisons between the entities involved. Tversky (1977) introduced an alternative way to measure similarity where the comparison between two entities is established by observing their common and distinctive features. Other approaches rate similarity based on transformational distances, i.e. similarity between two entities is inversely related to the number of operations required to transform one entity into the other (Hahn et al., 2003). It has been argued that similarity assessment is a more dynamic process, in which the representations of the entities under comparison and the choice of relevant properties "are flexible, change over time, and change in response to context" (Spencer-Smith and Goldstone, 1997, p. 54). Furthermore, Goldstone (1999) notes that different processes for assessing similarity are probably used for different tasks, domain, and stimuli. This observation assumes greater relevance as we move the discussion of similarity modelling into the music domain. Relations of similarity are omnipresent in music and can be found in early to contemporary works. Thus similarity has naturally been an important tool for the analytical study of music. One such example is paradigmatic analysis (Ruwet, 1972; Nattiez, 1975), where the relationships between parts of a musical work are established based on notions of repetition and similarity between musical fragments within the piece. Cambouropoulos points out that the application of the paradigmatic methodology to the analysis of melodies carries some "practical difficulties" mainly due to "complex issues such as the selection of important musical parameters for the description of musical entities, the hierarchic organisation of musical structure and the segmentation of a musical surface" (Cambouropoulos, 1998, p. 12). Thus, the literature seems to suggest that no matter what method is used to determine similarity ratings, they will depend considerably on the choice of features used to represent or observe the entities under comparison.

Melodic similarity and melodic segmentation are often bound together. The notions of identity and similarity are at the centre of the *General Computational Theory of Musical Structure* (Cambouropoulos, 1998), which aims to generate a structural description for a musical surface, by segmenting and categorising the segments found. Cambouropoulos writes that musical similarity "strongly affects the emergence of significant musical entities (e.g. motives) which in turn contribute towards a more integrated segmentation [of

a musical surface]" (Cambouropoulos, 1998, p. 32). Lerdhal and Jackendoff's *Generative Theory of Tonal Music* also highlights the structuring role of similarity and the importance of musical parallelism in the formation of groups, but the authors are unable to state precisely what conditions must be met by two passages before they can be seen as parallel: "when two passages are identical they certainly count as parallel but how different can they be before they are judged as no longer parallel" (Lerdahl and Jackendoff, 1983, p. 52).

According to Wiggins (1999) a "realistic analysis of melodic similarity is not a matter of comparing notes...it is a matter of comparing perceptions". Although well established that similarity is a key aspect in musical listening and understanding, it is not easy to underpin exactly what factors make listeners decide for such similarity judgements. Therefore, it is important to look at experiments in music listening, to be able to integrate perceptual or cognitive principles as heuristics for models of similarity.

There have been several empirical studies concerning the study of human similarity judgments in music, which we now discuss. Melodic similarity judgements seem to be affected by alterations or transformations that differentiate two melodies being compared. The tolerance of the listener to these alterations or transformations is therefore an important point of analysis. Rolland and Ganascia (1999) make a distinction between two types of differences that may affect the comparison of two melodic sequences. 'Local differences' refer to the absence, addition or modification of individual notes and can include changes in pitch, rhythm and loudness. Changes in pitch may occur in various dimensions such as pitch contour, pitch chroma and pitch height. 'Global differences' may include transformations such as transposition, inversion, reversal of entire phrases or melodies or even changes in tempo. Transposition of melodies may be further analysed with respect to direction, pitch distance and key distance.

Empirical research seems to support the idea that melodic contour is an preponderant cue to the listener when recognizing melodies and establishing similarity relations. In an early study by White (1960) subjects were asked to identify short melodic sequences which were transformations of excerpts of a set of well-known melodies. Most transformations involved some linear operations on the intervallic information of the melodies such as multiplications, additions or subtractions (absolute and relative) on the sizes of the intervals. Results revealed that the least disruptive transformations were the ones that preserved the relative magnitude of the intervals between notes and therefore preserved the pitch contour. In similar experiments, this time using transformed versions of unfamiliar isochronous melodic sequences, Massaro et al. (1980) showed that all contour

preserving transformations yielded the best judgments of similarity. On the other hand Hofmann-Engl and Parncutt (1998) argue that interval difference is a better predictor of similarity judgments than contour difference. Both interval difference and contour difference were analysed using multiple regression, with respect to the listeners' judgments, revealing that the contribution of contour was insignificant, and suggesting that contour difference was embedded in interval difference.

Experiments with infants have revealed that they respond to contour before they respond to melody, that is, they cannot distinguish between a song and a melodic alteration of that song, so long as contour is preserved (Lamont and Dibben, 2001). According to Hulse and Page (1988), only as the child matures is he able to attend to the melodic information. It has been argued that the contour of a melody might be more readily processed because it is a more general description of a melody, and therefore it subsumes interval information. It is only with increasing familiarity, or increasing cognitive abilities, that the intervallic details become perceptually important (Levitin, 1999).

Another widely discussed issue relating to pitch alterations is octave equivalence. Octave equivalence or octave generalisation refers to the perceived similarity of notes standing in an octave relationship, meaning that, for example, an ascending major ninth should be perceived as similar to an ascending major second. It has been shown that in a task of recognition, octave generalization is tolerated by listeners as long as it does not change the melodic contour, although most authors report that it increases the difficulty of melody recognition as it violates the relative height of the successive intervals (Deutsch, 1972; Massaro et al., 1980). Even simple well known melodies, when their pitches were randomly split into different octaves (preserving pitch chroma but altering pitch contour), they become almost unrecognisable (Dowling and Hollombe, 1977).

We now look briefly at some global transformations affecting melodic similarity judgements, starting with transposition. Research findings seem to convey that similarity decreases significantly with increasing transposition interval (Hofmann-Engl and Parncutt, 1998). Hofmann-Engl and Parncutt (1998) looked at the influence of the size of the transposed melodic fragments under comparison. They showed that smaller fragments would yeld lower mean similarity ratings, suggesting that listeners may adapt to a transposition after hearing a certain number of tones.

It has been claimed that transpositions to near keys (within the circle of fifths) are perceived as more similar to the original melody than transpositions to distant keys (Krumhansl, 1990). However Egmond et al. (1996) concluded that both pitch distance and key distance play a role in similarity, but no clear evidence was found that they

are interdependent. Using a multiple-regression analysis they concluded that the effect of key distance is minimal, pitch distance being the only factor significantly affecting the perceived similarity of the transposed melodies. The same study, also showed that direction of transposition appeared insignificant as a predictor of similarity.

A comparison between exact (i.e. preserving interval distances between notes) and inexact transpositions concluded that the latter are judged as less similar to the original melodies than the former (van Egmond and Povel, 1996). The same study was further extended by making a distinction between two types of inexact transpositions: diatonic and chromatic, but the authors point out that the differences found are not sufficient to make predictions concerning the size of the effects (Trainor and Trehub, 1992). It seems that although in specific situations transposition may have an effect on similarity judgements, recent studies indicate that infants, like adults, retain melodic information primarily in relative pitch form, and in a task of recognition, showed no preference for listening to a transposed version compared to the original pitch version of a familiar melody (Platinga and Trainor, 2005).

Other types of global transformations, whose detail is not relevant here, have been the subject of empirical studies, such as inversion (Dowling, 1972; Krumhansl, 1991; Hofmann-Engl and Parncutt, 1998) and reversal (White, 1960; Dowling, 1972).

The influence of time-based transformations in melodic similarity has, to the best of our knowledge, not been addressed as frequently or as explicitly as it has with pitch-based transformation. Perhaps this is because the study of rhythmic transformations in isolation is too abstract to shed light on similarity ratings of melodic sequences where pitch and rhythm are bound together.

Tempo, here seen as a time-based global transformation, has been looked at in some studies. Hofmann-Engl and Parncutt (1998) showed that tempo changes of factors from two to six did not have a significant effect on similarity judgments, suggesting that the listeners understood similarity as tempo invariant in context of isochronous fragments and therefore were able to separate tempo from other parameters. Changes in tempo are likely to affect more significantly complex melodic transformation such as transposition, inversion or reversal. This is because the perception of some of these transformations may involve memorisation and subsequent time-consuming mental processing over the melodic material heard. Indeed, the recognition of complex melodic transformations has been reported to suffer from faster playing of the melodic sequences (Dowling, 1972).

2.1.3 Memory in Music Listening

Memory is intrinsically related to the way we perceive the sensory world around us and the perception of music, as an auditory experience, is no exception. Several studies have shown how memory may influence listeners in the way they construct a mental schema of a musical work. In a real listening situation, music is presented progressively to the listener as the piece unfolds in time. Because of memory limitations, listeners are only capable of retaining some of the information on each hearing.

An important factor in auditory perception is echoic memory (Eysenck and Keane, 1995). Echoic memory allows auditory information to persist for some time after the end of stimulation. Studies have shown that this memory is limited in time (Davelaar et al., 2005) and the number of distinct elements that it may contain (Miller, 1956).

Short-term memory effects are also related to the notion of 'perceptual present' which corresponds to a period of time during which auditory stimuli are held present for perception, meaning that several past (although relatively recent) stimuli may draw the listener's attention, and may be retained as the actual most recent or most prominent stimuli.

Experiments on free recall [1] of items memorised in a sequence show that people have high recall for the first few items (primacy effect) and for the last few items (recency effect) but poor recall for the middle items (Ward and Lockhead, 1970). Primacy is usually attributed to favoured rehearsal of the first items perceived (longer time to transfer to long-term storage), while recency is due to persistency (last items have higher strength) in working memory or echoic memory.

Primacy effects have been observed in music listening. Deliège (1997) reports that "the impact of the first cues encountered seems paramount in shaping the imprint". Experiments have also provided evidence that recognition and comparison of melodic patterns is frequently made on the basis of the very beginning of the sequences heard (Deliège, 2001). This means that in some cases listeners may judge as similar, sequences which have similar beginning but distinct middle or ending sections. Recency effects in listening have also been reported by Krumhansl (1990) and Hofmann-Engl and Parncutt (1998).

It is likely that the detection of segmentation boundaries in the music will enhance the effects of recency and primacy, as these may then occur with respect to every distinctly perceived segment. Bigand (1993) has suggested that the listener perceives a

[1] recalling the items to be recalled in any order

musical surface by focusing on successive zones, that can be viewed as a "sliding window" along the musical piece. The limited size of this window (determined by memory restrictions) is a constraint in the establishment of temporal relations between the musical elements, a limitation also already acknowledged by Lerdahl and Jackendoff (1983, p. 21).

Other studies have reported the influence of sequential effects in rating or categorization procedures (Jestead et al., 1977; Ward and Lockhead, 1970; Brown et al., 2002). These findings, although not all particular to the musical domain, reveal the dynamics of some cognitive processes which are likely to be involved in the organisation of musical ideas, including the detection of segment boundaries, or the making of similarity judgments over the course of listening to a piece of music. Meyer (1973) referring to the evaluation of similarities between musical elements, points out that "the more regular and individual the pattern (and, of course, the more alike events are in interval, rhythm, etc.) the greater can be the temporal separation between model and variant and the greater the variety of intervening motifs, with the conformant relationships still recognisable". Thus, the similarity between two musical motives not only depends on the comparison between its musical features but also on how strongly these features can be remembered by the listener after hearing the first motif or motifs.

Memory retrieval depends on the similarity of the retrieval cue with stored memories, but factors like repetition and rehearsal also affect our ability to retain and remember musical information. The recollection of certain musical features or patterns may be observed due to either the occurrence of repeated similar patterns within the piece or by repeatedly hearing that same piece. Several empirical results account for the importance of repetition in the establishment of similarity relations and abstraction of thematic contents in music (Pollard-Gott, 1983; Krumhansl, 1990; Deliège, 1997). The effects of repetition can be conditioned by temporal or attentional factors. Dowling (1972) reports how the speed of presentation of melodic material affected recognition and similarity judgements but he remarks that "further work is needed to decide whether the advantage of slow presentation lies in the subjects' being better able to store the stimulus, or in the added time they have to manipulate the material [in memory]"

In summary, memory has a preponderant role in the way humans experience music, and has implications in the listening process, particularly when recognition and similarity judgements are involved. Without memory, subjects would not be able to perceive the similarities that relate distant sections of a piece of music, or that relate a musical passage to a family of musical works. It was shown that the capacity of listeners to

abstract musical contents is related to the exposure and familiarity with the musical material. Familiarity is therefore the result of a learning process, from the assimilation of a particular musical style, from previous memorisation of a musical reportoire or from recent hearing of a musical piece.

2.2 Learning and Expectancy in Music Perception

2.2.1 Meyer's Theories of Musical Expectancy

Leonard Meyer, in his book *Emotion and Meaning in Music* (Meyer, 1956), proposes that listeners' affective responses are the result of musical expectations. According to Meyer, inhibition or delay of tendencies or resolution during the course of listening generates uncertainty. He writes that "suspense is essentially a product of ignorance as to the future course of events" establishing the relationship between acquired knowledge and expectation (Meyer, 1956, p. 27).

Although Meyer acknowledges the relevance of the Gestalt principles in music perception, but he stresses that learning has a preponderant role in music understanding and musical expectations. He argues that it is the deviations from a sort of ideal Gestalt structure that allows expectations to be aroused and manipulated. As he notes, "the better the psychological organization, the less likely is it that expectation will be aroused." Meyer (1956, p. 87).

According to Meyer, musical experience is strongly influenced by past musical experiences and stylistic exposure. From this point of view, he writes, "what a musical stimulus or series of stimuli indicate and point to are not extramusical concepts and objects but other musical events which are about to happen. That is, one musical event (be it a tone, a phrase, or a whole section) has meaning because it points to and makes us exect another musical event." Meyer (1956, p. 35). This view corroborates the idea that musical experience is context-driven.

Meyer recognised the affinity between expectation and information theory (Meyer, 1967) and in particular with the notion of entropy. He explains that the acquisition of regularities in music facilitates learning, mostly because order and regularity make prediction possible: "the arrival of a predictable, regular event rewards the listener, [...] and thereby encourages and reinforces his learning. [...] Manifest irregularity or randomness, on the other hand, precludes predictability; and by weakening the listener's sense of control, discourages learning" (Meyer, 1967, p. 278).

2.2.2 The Implication-Realisation Model

The Implication-Realisation (I-R) Model (Narmour, 1992, 1990) proposes a theory of melodic perception, that is intended to account for the listeners' expectations and perceived changes in melodic continuation, as a melody unfolds.

In the I-R Model, expectancy is driven by both top-down and bottom-up processes of implication (expectation). According to Narmour (1992, p. 53), "bottom-up processing is the subconscious processing of the individual parametric syntactic primitives from the incoming signal, whereas top-down processing is the constructive matching of highly conformant (albeit schematic) style structures to the input events at hand".

Top-down processing is related to prior learning of musical structures including knowledge of a piece (intra-opus learning) or knowledge of the style of the piece (extra-opus learning). Bottom-up processing is said to operate in a mechanistic way, and is associated with innate unconscious processing of melodic primitives such as intervals, registral directions, durations, and melodic consonances/dissonances. Narmour argues that these two separate expectation systems are independent but interact and he writes that "any full explanation of musical phenomena must always attend to both since both simultaneously operate in the listener's apprehension of the real world" (Narmour, 1992, p. 36).

Narmour deals with melody mainly as a note-to-note phenomenon and describes the cognition of melodies as a succession of points of implication and points of closure. The last two tones in a melody at a point of implication form what he calls the *implicative interval*, and the next interval that follows in the melody is called the *realised interval* (Krumhansl, 1997). The I-R Model builds on Meyer's theories of melodic expectancy, and hypothesises that implications of continuation are subconsciously perceived by the listener mainly due to bottom-up Gestalt principles of similarity, proximity and good continuation. The model includes five principles that define how bottom-up expectancies are associated to implicative intervals. These principles are summarised by Krumhansl (1997) and can be described as follows:

registral direction specifies the implied direction of the realised interval relative to the direction of the implicative interval. This principle states that if the implicative interval is small then the registral direction is expected to continue (Gestalt principle of good continuation). If the implicative interval is large then the registral direction is expected to reverse (Gestalt principle of symmetry).

intervallic difference concerns the implied size of the realised interval relative to the size of the implicative interval. The principle states that if the implicative interval is small then a similarly sized realised interval is implied. If the implicative interval is large then a smaller realised interval is implied.

registral return governs the size of the interval between the first tone of the implicative interval and the second tone of the realised interval. This principle is satisfied when the second tone of the realised interval falls wihin two semitones of the first tone of the implicative interval

proximity governs the size of the interval between the second tone of the implicative interval and the second tone of the realised interval. This principle is satisfied when the second tone of the realised interval falls within 5 semitones of the second tone of the realised interval. (more simply, proximity holds if the realised interval is five semitones or less). Proximity is a graded variable ranging from a maximum corresponding to zero semitones (unison) to a minimum of five semitones (perfect fourth).

closure specifies the pairs of implicative and realised intervals that produce a sense of closure. Closure is strongest when registral direction reverses and a large implicative interval is followed by a smaller realised interval. Closure is moderately strong when either registral direction reverses or a large implicative interval is followed by a smaller realised interval. When neither of these conditions is present then the interval is considered open [2].

In the formulation of the I-R model, more focus is given to intervallic and registral parameters, while other aspects such as rhythm, meter and harmony, which according to Narmour may reinforce or inhibit melodic expectations, seem to lack the formality and depth of the rest of the theory (Cross, 1995).

The I-R Model has been the subject of several experimental studies which have empirically verified some of its principles (Cuddy and Lunney, 1995; Russo and Cuddy, 1996; Krumhansl, 1997). Narmour concentrates mostly on a lower analytical level of note-to-note relations and seems to attribute less relevance to higher-level melodic structural analysis. Some empirical studies, however, have correlated Narmour's principles to more macro-analytical judgements of listeners, such as melodic cohesiveness and

[2]This is assuming that other factors such as metrical or durational stress or resolution of dissonance, are absent

pleasingness (Russo and Cuddy, 1996), suggesting that the application of the I-R Model may extend beyond the description of note-to-note expectancies.

Narmour highlights the musicological applicability of his theory and suggests that it may be used to generate a representation of how listeners might apprehend and encode melodies. From this perspective, the I-R model offers an analytical system where expectations induce points of closure or continuation, and these could arguably be used as structural descriptors, identifying points of change in the melodies.

Although complex in its formulation, the I-R Model provides a reflection on the dynamic role of expectation in music listening. In particular, it argues for a separation between contributing learned expectations and innate bottom-up Gestalt-based expectations. This assumption of the theory has been challenged both theoretically (Cross, 1995) and experimentally (Pearce and Wiggins, 2004) and is also addressed in the course of the present work.

2.2.3 Cue Abstraction and Imprint Formation Theory

Irène Deliège's (1997) psychological theory of Cue Abstraction and Imprint Formation brings new insights to explain how listeners abstract a mental representation of a given musical work. The theory maintains that prototypical descriptions of a musical piece (the imprints) are acquired and held in memory by subjects, during the course of listening, on the basis of the repetition and salience of small musical cells (the cues).

The cues (also referred to by the author as *primary cells*) are short patterns which become quickly fixed in memory because of their salience, their pertinence and repetition either in a literal or in a varied form (Deliège, 1997). Cue abstraction can be viewed as a process of perceptual reduction. The cues contain the invariants of the musical discourse and their function is to provide abbreviations of longer sequences: "cues are the starting point for processes of comparison between old and new entries in working memory, on the basis of principles of sameness and difference" (Deliège and Melén, 1997). This idea has also been supported by (Levitin, 1999) by suggesting that melodies often carry "flags" or "markers" that act as starting points for different sections in a song.

The imprints are memory recollections which result from the accumulation of varied repetitions of cues, "a sort of 'résumé' of the main coordinates of a set of presentations around the same basic structure" (Deliège and Melén, 1997). The idea that imprints are formed based on accumulation and repetition is a key aspect in Deliège's theory, as it highlights the importance of memory in music listening.

A vast number of empirical results (Deliège and Melén, 1997; Deliège, 1998, 2001) supports the idea that cue abstraction takes place while listening to music, although the exact definition of musical cue seems yet unclear in many aspects. Are the perceived cues constructed by the listener or are they found in the musical surface?

Deliège suggests that the types of cues that may be perceived in a musical piece are dependent on the cultural and historical provenance of the piece. Does this mean that cue abstraction can only take place if the listener is familiar with the music? If the listener is not familiar with a particular musical system what kind of processes does she use to abstract cues from the music? Krumhansl (1991) has described the capacity of listeners to abstract and recognize music written in an unfamiliar style. However other experiments have shown that musically trained listeners are able to extract thematic similarities from tonal pieces more easily than from atonal pieces, provided listeners have previously heard the pieces in both cases (Lamont and Dibben, 1997).

The cue abstraction process has been observed in different experiments performed with different musical reportoires (including atonal pieces) and Deliège and Melén (1997) claim that the generality of cue-abstraction lies in the fact that that process is accompanied by the formation of the imprint, which embodies the stylistic characteristics of the work in the course of listening. We argue that this assumption may only be partially valid. Deliège's theory advocates a particular form of listening, which is somehow implied by the type of experiments she has performed to support it. For some musical works, familiarity with the genre may not be enough to facilitate imprint formation. For example, if the music does not contain any motivic/thematic recurrence, it may not accommodate prototypical descriptions of its contents.

Deliège has observed that musicians and non-musicians perform differently in their ability to abstract musical contents, but that some level of cue abstraction always takes place, independently of the subject's musical training.

Although extensive empirical data seems to support the overall assumptions of Deliège's theory, but the latter fails to reveal in more detail the cognitive processes to which she alludes, leaving open the possibility for further research and experimentation.

2.3 Existing Models of Music Segmentation

2.3.1 Generative Theory of Tonal Music

The Generative Theory of Tonal Music (GTTM) by Lerdahl and Jackendoff (Lerdahl and Jackendoff, 1983) has been a recurring focus of debate in the music research community. The following overview of GTTM follows closely many of the original descriptions and definitions contained in the original publication (Lerdahl and Jackendoff, 1983), unless where explicitly referenced.

The theory is a top–down approach to the generation of a structural description for a musical work. Lerdahl and Jackendoff formalise a set of rules, which they claim, correspond to the intuitions of an experienced listener in the Western tonal idiom. The theory is greatly influenced by Chomskian linguistics and also incorporates concepts from Gestalt psychology.

The theory comprises four hierarchical components:

Grouping structure expresses a hierarchical segmentation of the piece into motives, phrases and sections.

Metrical structure expresses the intuition that the events of the piece are related to a regular alternation of strong and weak beats at a number of hierarchical levels.

Time-span reduction assigns to the pitches of the piece a hierarchy of "structural importance" with respect to their position and grouping and metrical structure. The authors define a time-span as "an interval of time beginning at a beat of the metrical structure and extending up to, but not including, another beat". A Time-span reduction consists of several levels of a simplified notation, combined in a tree-like structure. A branch in the time-span reduction tree indicates subordination.

Prolongational reduction assigns to the pitches a hierarchy that expresses harmonic and melodic tension and relaxation, continuity and progression.

The authors acknowledge that GTTM is restricted only to those components of musical intuition that are hierarchical in nature. Although, they claim, the theory takes into account the influence of non-hierarchical dimensions of musical structure, such as timbre, dynamics or motivic-thematic processes, these are not formally represented.

The starting point for the application of this theory to a musical piece is a representation containing a sequence of pitches and durations. Each one of the four hierarchical

structures is then generated by the application of two kinds of rules: 'well-formedness and 'preference rules'. Well-formedness rules specify the possible structural descriptions and are comparable to the rules of a linguistic grammar. They ensure the formal soundness of the structural descriptions. For example:

Grouping Well-Formedness Rule 4 If a group G_1 contains part of a group G_2, then it must contain all of G_2.

Preference rules designate, out of the possible descriptions, those that correspond to experienced listeners' hearings. As an example:

Grouping Preference Rule 6 (Parallelism) Where two or more segments of music can be construed as parallel, they preferably form parallel parts of groups.

Preference rules have no counterpart in a standard linguistic grammar, but were included because, the authors argue, "music contains more ambiguity than language and is not tied down to specific meanings and functions, as language is".

The GTTM relies on the general assumption that an important part of music understanding relies on the segmentation of a piece into constituent units such as motives, phrases, sections, etc. A common criticism of this theory is that it is restricted to tonal music, although some of the components of the theory, particularly the ones related to grouping structures, rely on more general perceptual principles. Some of the few experimental studies on the GTTM (Deliège, 1987; Bigand et al., 1994) show that from all four components of the theory (grouping structure, metrical structure, time–span reductions and prolongational reductions), only the grouping rules seem to be empirically supported. Bigand (1993) also draws attention to the fact that GTTM is a model of comprehension of tonal music at the final stages of cognitive processing, and therefore it does not take into account cognitive processes which occur during real-time listening.

Other criticisms of the theory address the fact that the theory lies somewhere between a grammar and a rule-based analytical method. The inclusion of preference rules raises the problem of quantification of rule strengths, so crucial for the implementation of a theory. Also, some of these rules are fairly vague in the way they are to be applied. For example, regarding preference rule GPR 6 on parallelism, described previously, the theory does not provide any explanation of how two judge two music segments as parallel. The authors of GTTM acknowledge that "the problem of parallelism, is not at all specific to music theory; it seems to be a special case of the much more general problem of how people recognise similarities of any sort" (Lerdahl and Jackendoff, 1983, p. 53).

2.3.2 Local Boundary Detection Model

The Local Boundary Detection Model (LBDM) is a segmentation algorithm that was developed initially as part of Cambouropoulos' General Computational Theory of Musical Structure (GCTMS) (Cambouropoulos, 1998): a style–independent procedure to generate a structural description for a musical surface.

The LBDM calculates a boundary profile for a melody, using Gestalt-based identity–change and proximity–difference rules, applied to several parameters describing a melody. Segmentation is then achieved by detecting points of maximal change, that the author claims, are used by the listener to identify local boundaries in a melody (Cambouropoulos, 1998, p. 114).

According to a refined version of this algorithm (Cambouropoulos, 2001a), LBDM takes as input a melodic sequence converted into several independent parametric interval profiles $P_k = [x_1, x_2, ...x_n]$ where $k \in \{pitch, ioi, rest\}, x_i \geq 0$ and $i \in \{1, 2, ...n\}$. A *Change* rule assigns boundaries to intervals with strength proportional to the degree of change between neighboring consecutive interval pairs. Then a *Proximity* rule scales the previous boundaries proportionally to the size of the intervals.

The strength of the boundaries at each interval x_i is given by

$$s_i = x_i \times (r_{i-1,i} + r_{i,i+1}) \tag{2.1}$$

where

$$r_{i,i+1} = \begin{cases} \frac{|x_i - x_{i+1}|}{x_i + x_{i+1}} & x_i + x_{i+1} \neq 0 \wedge x_i, x_{i+1} \geq 0 \\ 0 & x_i = x_{i+1} = 0 \end{cases}$$

For each parameter k a sequence s_k is calculated, then all sequences are normalised and combined in a weighted sum to give the overall boundary strength profile. The suggested weights for the 3 different parameters are $w_{pitch} = w_{rest} = 0.25$ and $w_{ioi} = 0.5$ (see Thom et al. 2002 for an overview on the behavior of the LBDM with different parameter tunings). The local peaks in the resulting boundary profile indicate local boundaries in the melodic sequence. A threshold must be defined a priori, above which, a peak is identified as a boundary. For additional details on the implementation of the LBDM the reader is referred to Cambouropoulos (2001a).

The LBDM indicates several possible segmentation boundaries, but it does not have the ability to determine which boundaries should be considered as the most significant segmentation markers, allowing these decisions to be made by subsequent pattern de-

tection and selection procedures. As mentioned previously, the LBDM is not a complete model of grouping in itself, as it relies on complementary models (i.e. pattern similarity) to select the most relevant boundaries. Although in that context this may not be considered a weakness of the model, excessive boundary generation may become a disadvantage if we intend to use the LBDM in isolation, and when segmentation is to be used as a reliable data reduction technique.

The LBDM has a fairly short memory as it considers at most 4 consecutive events at a time. As a consequence, there is limited interaction between neighboring boundaries and sometimes small "oscillations" can be identified as salient boundaries. This type of limitation has also been referred to by Lerdahl and Jackendoff (1983) in their Generative Theory of Tonal Music.

Cambouropoulos has used his computational model to replicate some of Deliège's experimental procedures, establishing a parallel between his model and the theory of cue abstraction and imprint formation (Cambouropoulos, 2001b). In his model, similarity matching between melodic segments (extracted in the segmentation phase) is performed according to a predefined set of properties of the segments, such as diatonic pitch, contour information, duration patterns, and other statistical attributes. Musical cues are therefore viewed as the prominence of these properties in the melodies. The thresholds used to rate the properties of the segments "were determined in an ad hoc manner" and Cambouropoulos (2001b) recognises that further research would be necessary to establish a more cognitively pertinent and computationally useful description of such melodic properties.

Cambouropoulos also establishes a parallel between imprints and the categories of melodic segments. The categories are sets of segments, grouped on the basis of similarity comparisons, according to the different features that characterise each segment. He then describes a procedure to compute a prototype for each one of the categories found. Each prototype highlights the features that are most characteristic of the melodic motives in each category.

2.3.3 Grouper

Grouper is a computational implementation of a preference rule system for melodic phrase structure, and is one of the components of Melisma Musical Analyser developed by Temperley and Sleator[3]. Most components of this system, including Grouper, fol-

[3]An implementation and documentation of this system is available at: http://www.links.cs.cmu.edu/music-analysis/

low closely the preference rule framework described in detail in Temperley's book "The Cognition of Basic Melodic Structures" (Temperley, 2001).

Grouper then searches for all possible 'well-formed' grouping analyse, that best satisfy the set of preference rules. Temperley describes a 'well-formed' grouping analysis as simply some segmentation of the input melody into phrases, where every note (of the melody) must be contained in a phrase, and the phrases must be non-overlapping (Temperley, 2001).

The input to the model is a MIDI-like 'note list' (or 'piano-roll') that includes the onset-time, offset-time and pitch of a series of notes. The model also requires as input the metrical information. This is provided to the model as a series of beats aligned with the 'piano-roll'. Each beat has a time point and a level number, indicating the highest metrical level at which that time point is a beat.

Grouper uses 3 preference rules in a similar fashion to the *General Theory of Tonal Music* (Lerdahl and Jackendoff, 1983), and can be summarised as follows:

PSPR 1 (Gap Rule) Prefer to locate phrase boundaries at a) large IOIs (inter-onset intervals) and b) large OOIs (offset-to-onset intervals). A gap score is obtained by weighting the sum of both a) e b).

PSPR 2 (Phrase Length Rule) Prefer phrases that have roughly 8 notes in length. The application of this rule penalises only slightly phrases whose length is close to 8, but penalises highly those that considerably deviate from this length.

PSPR 3 (Metrical Parallelism Rule) Prefer to begin successive groups at parallel points in the metrical structure.

Most of the experimental work done with Grouper (Temperley, 2001; Thom et al., 2002), used quantised 'piano-roll' input data, that was generated from scores (given the tempo and notated durations). Temperley observes that quantised note representation may differ significantly from one obtained from a live performance. Live performance data may contain offset–to–onset gaps and tempo nuances that can be important cues to phrase structure (Temperley, 2001).

The Phrase Length Rule follows the assumption that phrase length must be on average 8 notes. This rule derives from the fact that, as the author argues, "phrases are rather consistent in terms of their number of note" (Temperley, 2001). This value was determined empirically from a set of melodies, used to find the optimal values for all parameters. Although the author does not provide an objective explanation for the value of this

parameter, he suggests that it may be a result of constraints on performance: "phrases normally correspond to vocal breaths, and it is difficult to sing more than a certain number of notes in one breath"; or a result of information-processing contraints in perception: a clear reference to Miller's 7+/-2 number of elements that can be chunked together as a perceptual unit (Miller, 1956).

The system only derives one level of grouping, referred to as the 'level of phrase'. Temperley argues that this seems to be the level that is clearest and least ambiguous in perception.

Because metrical information is necessary as input in Grouper, the Melisma Musical Analyser provides a module that can generate a beat list from the original note file. However, Grouper is limited to input melodies that have perfectly regular metrical structure. The performance of Grouper (and similarly for all other modules in Melisma) depends on a series of parameters that can be adjusted for 'optimal' performance. These include weights that control the placement of phrase boundaries according to criteria such as difference from optimal phrase length, and possible gaps between phrase boundaries. Other parameters define penalties for boundaries that are out-of-phase with different existing beat levels. See Thom et al. (2002) for a discussion of the influence of some of these parameters.

Grouper is part of a wider computational model of music cognition. Temperley evaluates the results of his model by comparing its output with the analyses of music scores. For many short western melodies (e.g. folk songs) this might be acceptable since in many cases, listeners perceptions can be predictable. But in the general case, to make claims about how listeners identify melodic phrases, one would expect that some real listening data should be matched with the results of the model.

2.4 Summary

This chapter provided a review on perceptual and cognitive factors that play a role in melodic perception. A particular focus was made on Gestalt theories and similarity, which are recurrently associated with melodic segmentation. An overview of several empirical studies on music perception and cognition has shown that listeners respond differently to different melodic attributes and melodic transformations when abstracting musical contents. This findings will inform the choices made for melodic representation for the segmentation models we propose to implement.

We reported evidence on the role learning in music perception, in particular, on how musical expectancy can influence the perception of boundaries in melodies. These findings constitute the main motivation for two memory-based models of melodic segmentation developed in the course of this research. A review on exiting segmentation models presented in the last section of this chapter, reveals that memory effects are often not taken into account, while exiting memory-based approaches to music learning have focus often on music composition and rarely on the abstraction of musical structure.

In the next chapter we provide a more specialised overview on information theory and probabilistic modelling, to provide the background for the second of the two segmentation models to be developed.

Chapter 3

Information Theory and Probabilistic Modelling

This chapter provides some background on Information Theory, introducing notions such as information and entropy which constitute the theoretical support to our melodic segment prediction method. What follows is an overview of probabilistic language models with particular emphasis on Mixed-Memory Markov models, which were adopted to implement our probabilistic model for melodic segmentation.

3.1 Information Theory

The fundamentals of Communication Theory or Information Theory (IT) are attributed to Claude Shannon (1948), who systematised the basic components of a communication system: the information source (transmitter), the communication channel, and the receiver. Shannon addressed the problem of the transmission of a message with maximum accuracy and efficiency over a noisy channel, establishing quantitative ways of measuring the information contents in a message being transmitted, received, or stored.

Information theory often describes the properties of the source in terms of probabilities. Typically a source of information will have an alphabet of N possible symbols $[s_1 \cdots s_N]$ and each symbol is associated with a probability of ocurrence $[p_1 \cdots p_N]$. A *message* is then a sequence of symbols drawn from the alphabet according to the corresponding probabilities.

Because the source output is not known to the receiver in advance, it can be said that *information* is how much more we know after each new symbol is seen by the receiver, than we knew before. Intuitively, the least probable symbols convey most information. *Information* is therefore also a measure of uncertainty of a message. In other words, the more uncertain or unpredictable a message is, the more information it conveys.

Information or self-information The information content of a symbol x_i with probability $P(X_i)$ is given by

$$I(x_i) = -\log(P(x_i)) \tag{3.1}$$

If the base of the logarithm is 2 then information is measured in *bits*. The use of the logarithm is convenient because it complies with essential properties of self-information. For example, given the occurrence of two successive independent symbols with probabilities $P(x_1)$ and $P(x_2)$, the information content of the sequence is $I(x_1, x_2) = -\log(P(x_1)P(x_2)) = -\log(P(x_1)) - \log(P(x_2)) = I(x_1) + I(x_2)$.

It may be useful to measure the typical source message information content, rather than that of individual symbols. The definition of self-information can be extended to quantify the average information conveyed by a source, leading to Shannon's notion of *source entropy*:

Entropy Given a source of alphabet size N with symbol probabilities p_i, *source entropy*, or average information content per symbol, is given by:

$$H(X) = \sum_{i=1}^{N} p_i I(x_i) = -\sum_{i=1}^{N} p_i \log(p_i) \tag{3.2}$$

As with the definition of *self-information*, if we adopt a base 2 logarithm, then *entropy* is expressed in $bits/symbol$. Because p_i are probabilities, it follows from Equation 3.2 that $H(X) \geqslant 0$ for all X. $H(X) = 0$ only when $P(X) = 1$, that is, when the value of X is pre-determined no new information is conveyed.

To illustrate both concepts of *information* and *source entropy*, a simple example is presented in Figure 3.1, where we consider an information source that produces sequences of symbols, based on the toss of a coin. The example considers two distinct situations corresponding to the toss of a fair coin and a biased coin. In the case of the fair coin both outcomes ($heads$ and $tails$) are equiprobable and thus both convey 1 bit of information. In the case of the biased coin, the less probable occurrence of $tails$ conveys 2.32 bits of information, considerably more than the 0.32 bits for $heads$, although the average information of this source (0.72 bits/symbol) is less than 1 bit/symbol for the equiprobable case.

When an information source has a finite alphabet of symbols whose likelihood of occurrence is statistically independent, it is designated as a *discrete memory-less source*, also referred to by Shannon as a zero-order process. So for a source with an alphabet

Fair coin	$P(h) = 0.5$ $P(t) = 0.5$	$I(h) = I(t) = 1$	$H(X) = 0.5 * 1 + 0.5 * 1 =$ $= 1$ bit/symbol
Biased coin	$P(h) = 0.8$ $P(t) = 0.2$	$I(h) = 0.32$ $I(t) = 2.32$	$H(X) = 0.8 * 0.32 + 0.2 * 2.32 =$ $= 0.72$ bits/symbol

Figure 3.1: Comparing *information* and *souce entropy* for a process based on a fair coin toss and a biased coin toss

of N equiprobable symbols, we know that $P(X) = \frac{1}{N}$, for all X. So from Equation 3.2 we have, $H(X) = -\sum_{i=1}^{N} \frac{1}{N} \log(\frac{1}{N}) = \log_2 N$ bits/symbol. More generally, for any memory-less source, with N possible symbols, we can establish lower and upper bounds for H as $0 \leqslant H \leqslant \log N$.

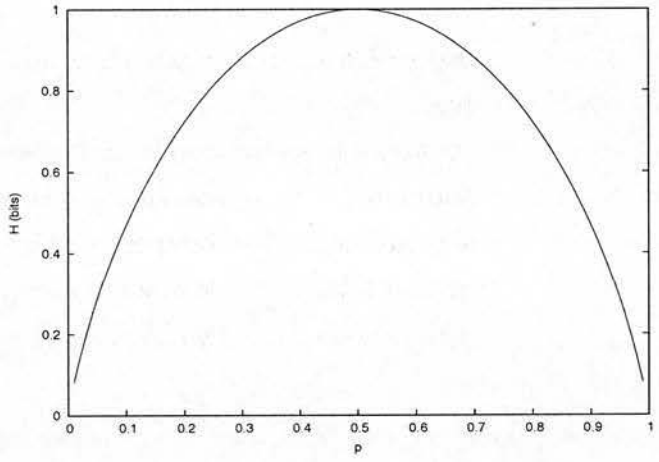


Figure 3.2: Entropy for a binary memory-less source, where two possible outcomes have probabilities p and $1 - p$

We have seen an example where the successive symbols at the source are independent. However this is not true for many temporal sequences of symbols. It can happen that the occurrence of some symbols may depend to some extent on the occurrence of the previous ones. In other words, messages may exhibit statistically detectable structural regularities which may become apparent after a certain period of time or when enough messages have been received and analysed. The receiver is then able to make predictions about the source, by inferring not only the vocabulary of the source and frequency of occurrence of symbols, but also the inherent structural properties of the messages. Shannon provides an example for the English language, where sequences are generated

based on the acquired statistics and dependencies of a 27 symbol alphabet (26 letters and a space):

1. Zero-order approximation (symbols independent and equiprobable).
   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.
2. First-order approximation (symbols independent but with frequencies of English text).
   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.
3. Second-order approximation (bigram structure: symbols depend on previous letter).
   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.
4. Third-order approximation (trigram structure: symbols depend on two previous letters). IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

(Shannon, 1948, pg. 7)

Shannon further expands this idea by considering word units rather than letters:

1. First-order word approximation (words are chosen independently but with frequencies of English text).
   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TOOF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.
2. Second-order word approximation (the word transition probabilities)
   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

(Shannon, 1948, pg. 7)

These examples, albeit simple, illustrate the potential of statistical processes, but also highlight the dependence of such processes on context size, the level of representation, and the choice of structural dependencies used for probabilistic analysis.

Mutual Information The (average) *mutual information* between two random variables X and Y is the quantity

$$I(X;Y) = H(X) - H(X|Y) \tag{3.3}$$

where $H(X)$ is a measure of uncertainty about X and $H(X|Y)$ a measure of uncertainty about X provided we know Y. Mutual information is given by the difference of these two uncertainties, and represents the decrease in uncertainty that results from the knowledge of Y. Note that $I(X;Y) \geqslant 0$ and equality only happens if and only if X and Y are statistically independent. Other properties of mutual information include:

$$I(X;Y) = I(Y;X) \tag{3.4}$$

$$I(X;Y) = H(X) + H(Y) - H(XY) \tag{3.5}$$

3.2 Probabilistic (Language) Modelling

Probabilistic models provide us with ways of analysing a set of data according to some unknown probability distribution. Probabilistic language models are useful because they allow us to make predictions about sequential combinations of words, for example determining the next possible outcomes in an incomplete sequence.

For the past several years probabilistic language models have been used for a variety of text and language processing tasks in areas such as natural language processing (Manning and Schüttze, 1999) and speech recognition (Rabiner, 1989).

3.2.1 Conditional Probability and the Notion of Context

Statistical models are not just based on raw frequency counts. The notion of context is central to probabilistic modelling. For a given distribution, we can say, at any given moment, that there is a known context which might condition the outcome of successive events. The simpler way to express this type of dependency in probabilistic terms is with conditional probabilities.

Conditional probability of an event A given the occurrence of an event B is defined as:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} \tag{3.6}$$

The value $P(A|B)$ is often referred to as the *posterior probability* of A since it is determined after the occurrence of B, whereas $P(A)$, referred to as the *prior probability* of A, is obtained in the absence of any context. If events A and B are independent then we know that $P(A \bigcap B) = P(A)P(B)$ so we get $P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$, meaning that no information is gained about A by knowing the occurrence of B.

More generally, given a sequence of symbols $W = (w_1, w_2, w_3, ..., w_n)$ we can determine its probability $P(W)$, as:

$$\begin{aligned} P(W) &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, ..., w_{n-1}) = \\ &= \prod_{i=1}^{n} P(w_i|w_1...w_{i-1}) \end{aligned} \tag{3.7}$$

3.2.2 N-Gram Models

It was shown previously that temporal sequences of events often have dependencies. N-gram models, a class of Markov models (also known as *Markov chains*) (Manning and Schüttze, 1999), are usually constructed from the statistical co-occurrences of adjacent symbols found in a given set of symbolic data (usually referred to as the *training corpus*). N-gram models can be used to determine the probabilities of sequences of symbols and are commonly derived from training sets of data that share the characteristics of the expected target sequences (or input) to be analysed.

An nth order n-gram model assumes that the probability of occurrence of a symbol depends on the occurrence of the previous $n-1$ symbols (equivalent to $(n-1)$th order Markov approximation). Hence the occurrence of a symbol is predicted according to a conditional distribution based on a limited context. For example, a bigram model (equivalent to a first-order Markov approximation) can be represented by a matrix containing the transition probabilities between adjacent symbols:

$$a_{ij} = P(X_{t+1} = s_j|X_t = s_i) \tag{3.8}$$

where $a_{ij} \geqslant 0, \forall i, j$ and $\sum_{j=1}^{N} a_{ij} = 1, \forall i$, and $S = \{s_1, s_2, ..., s_N\}$ is a finite set of symbols often called the *state space*.

Although in general we would like the order of the model to be high to allow for larger contexts, there are problems that result from large n. For higher order models, because training data is usually limited, some n-grams will have very low or even zero frequency, a problem usually referred to as 'data sparseness'. It is clear from Equation 3.8 that such models will result in poor probability estimations.

Several methods, usually referred to as smoothing methods, have been proposed (Manning and Schüttze, 1999; Chen and Goodman, 1996) to overcome the data sparseness problem. The aim of these smoothing methods is to derive good probability estimates based only on the available observed data. Some smoothing methods work by

readjusting the probabilities of the sequences, in order to guarantee that they all have a non-null frequency assigned to them. Other approaches such as linear interpolation work by expressing models of higher order as a weighted sum of models of lower order, which are less affected by data sparseness.

With linear interpolation the probabilities of a sequence of length l can be estimated by a weighted sum of n-gram probabilities from models of order $n <= l$. For instance, the probability of a trigram is determined by the weighted sum of corresponding unigram, bigram and trigram probabilities,

$$P(w_k|w_{k-3}, w_{k-2}, w_{k-1}) = \lambda_1 P(w_k) + \lambda_2 P(w_k|w_{k-1}) + \lambda_3 P(w_k|w_{k-2}, w_{k-1}) \quad (3.9)$$

where $0 \leqslant \lambda_i \leqslant 1$ and $\sum_i \lambda_i = 1$.

This technique is usually referred to as 'n-gram smoothing' (Chen and Goodman, 1996). The weights in Equation 3.9 can be set according to the application or can be inferred automatically from the training data using a maximum likelihood estimation method such as the Expectation-Maximisation Algorithm (Dempster et al., 1977)(Ney et al., 1994). For an overview on smoothing methods the reader is referred to Jelinek and Mercer (1980); Chen and Goodman (1996).

3.2.3 Mixed-Memory Markov Models

Mixed-memory Markov Models (MMM) provide a representation of higher-order models by combining several lower order models (Saul and Pereira, 1997; Saul and Jordan, 1999). Thus an nth order model over a discrete random variable W (with k possible values) can be expressed as:

$$P(w_i|w_{i-1}, ..., w_{i-n}) = \sum_{m=1}^{n} \phi_m a^m(w_i|w_{i-m}) \quad (3.10)$$

where $a^m(w_i|w_{i-m})$ is a $k \times k$ transition matrix containing the probabilities of the occurrence of a symbol at position i given the occurrence of a symbol at position $i - m$, and ϕ_m are the weighting coefficients of the mixture model. Ney et al. (1994) refer to these coefficients as 'memory weights' as they define the symbol-distance dependent weight of the influence of each predecessor symbol on the present symbol.

Equation 3.10 must also satisfy the following:

$$\sum_{i'} a^m(i'|i) = 1, \forall i, m \tag{3.11}$$

$$\phi_m \geqslant 0, \sum_m \phi_m = 1 \tag{3.12}$$

The coefficients of the mixture model can be estimated iteratively by an Expectation-Maximisation (EM) procedure (Dempster et al., 1977). For detailed description of this procedure the reader is referred to Ney et al. (1994).

The order of the model n determines the number of bigram models that are to be combined. This results in a model that grows linearly in the size of the context window used to determine each prediction. The MMM in equation 3.10 can be specified in $O(nk^2)$ parameters, in comparison with the $O(k^{n+1})$ parameters for the full-memory n-gram model in Equation 3.9.

It is important to observe the distinction between the mixture model in Equation 3.10 and models that approximate higher-order Markov models using a linear combination of nth order transition matrices, as defined in Equation 3.9. The MMM approximates a higher order model by taking a linear combination of non-adjacent bigram models.

When the size of the training corpus is small in comparison with the size of the vocabulary (state space) data sparseness can be reduced by using a mixture of bigram models. This characteristic of mixed-memory Markov models has been supported experimentally, showing that contrary to standard n-gram models, the number of unseen symbol combinations decreases with the order of the model (Saul and Pereira, 1997). This results from the fact that MMMs assign finite probabilities to any sequence of symbols $w_1 w_2 ... w_n$ for which any of the m-separated bigrams $w_m w_n$ are observed in the training set.

Although mixed-models cannot capture the full structure of equivalent n-gram models, there are situations where they can be advantageous. Because in MMMs only pairwise dependencies are stored, additional redundancy emerges from the data, and dependencies between distant symbols may be acquired. This is illustrated in the example in Figure 3.3 where the 2-separated bigram aa, registers three occurrences in the MMM but fails to be captured by the equivalent trigram model.

As a result, identical sequences may be assigned similar probabilities, so MMMs have the potential to represent some form of approximate similarity between symbol sequences, in probabilistic terms. Of course, the success of mixed-order models will

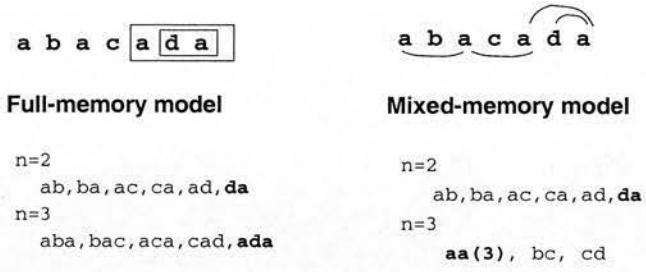


Figure 3.3: Comparison between n-gram and Mixed-memory models on the acquisition of symbol dependencies.

depend partly on the ability to capture these types of dependencies from the data.

As it was seen, MMMs are usually less affected by the complexity problem usually associated with higher-order n-gram models thus allowing an increase in the number of look-back dependencies that can be accounted for. But the effective order of the model depends mostly on the mixing coefficients c_m, which will determine the order equivalence of the mixed approximation (Saul and Jordan, 1999).

Chapter 4

Preliminary Experiments on Melodic Segmentation

In this chapter we present some preliminary work towards a memory-based model for melodic segmentation. We develop a new model for melodic segmentation based on the notion of melodic density, emphasising the role of short-term memory and time in music listening, by modelling the effects of recency in the perception of boundaries. We describe the model in detail and report some experimental results, by comparing it with the Local Boundary Detection Model (LBDM) (Cambouropoulos, 2001a), for a set of melody examples. Results suggest that this new model is more selective, as it generates fewer total boundaries but preserves most boundaries that coincide with the limits of melodic patterns. We discuss the limitations of modelling melodic segmentation based only on Gestalt-based grouping rules, and propose further developments.

4.1 Introduction

Melodic similarity, as mentioned in Chapter 2, plays a significant role in the perception of segment boundaries. Local segmentation methods based on Gestalt-based grouping rules alone don't always succeed in finding all relevant structural boundaries of a piece. However these segmentation methods, due to their simplicity, can be used successfully as a pre-processing stage for other tasks such as pattern discovery or music search.

Pattern finding algorithms, in particular, are known to be computationally expensive, and therefore can benefit from a reduction of the initial search space. A local segmentation can provide an efficiency gain by pre-processing a melodic sequence, and generating an initial set of boundaries which may be used as markers for pattern search (Cambouropoulos, 1998). One such method is the Local Boundary Detection Model

(LBDM) (Cambouropoulos, 2001a), a segmentation model that identifies discontinuities in a melodic surface based on Gestalt principles of perception.

The LBDM is an essential reference amongst segmentation algorithms, mostly due to its simplicity and generality (Cambouropoulos, 2001b,a). As the author emphasises, the LBDM is not a complete model of grouping in itself, as it relies on complementary methods (i.e. pattern similarity) to select the most relevant segment boundaries. The LBDM has a tendency to over-segment, when matched with the boundaries of larger recurring patterns, even if in some cases, excessive boundaries may correspond to smaller sub-patterns. Although in conjunction with other complementary modules this may not be considered a weakness of the model, excessive boundary generation may become a disadvantage if we intend to use the LBDM in isolation, and when segmentation is to be used as a reliable data reduction technique.

The LBDM has a fairly short memory as it considers at most four consecutive events at a time. As a consequence, there is limited interaction between neighbouring boundaries and sometimes small pitch or rhythmic oscillations can be identified as salient boundaries. This type of limitation has also been referred to by Lerdahl and Jackendoff (1983) in their Generative Theory of Tonal Music .

The short memory window of the LBDM makes it possible to segment a new sequence of notes without having to look at the whole sequence, prior to segmentation. This characteristic is desirable if we're interested in performing some form of real-time segmentation. On the other hand, if we are able to look at the whole sequence, we may be able to ponder the strength of a boundary according to the strength of neighbouring boundaries or even all other existing boundaries in the piece.

Research on auditory perception and memory has underlined the influence of time in the perception of differences and in the establishment of temporal relations in sequential processes. Studies have shown that listeners retain auditory information for some time, even after the end of stimulation (Eysenck and Keane, 1995). This means that several past (although relatively recent) stimuli may draw the listener's attention, and may be retained as the actual most recent and prominent stimuli. Some researches have suggested that listeners perceive a musical surface by focusing on successive zones, that can be viewed as a "sliding window" along the musical piece (Bigand, 1993). The size of this window (determined by short-term memory limitations) should restrict the amount of musical material that can be looked back on when processing a melodic sequence. Within this time window, recency effects are likely to apply, as documented by Krumhansl (1990) and Hofmann-Engl and Parncutt (1998).

The following section describes a new model for local segmentation. This model is an attempt to address some of these memory-related issues and follow on some of the ideas suggested from the analysis of the results of the LBDM. The model is also based on grouping rules but extends the idea of a memory window (as a context of events) by relating it to real-time issues such as recency and tempo.

4.2 Melodic Density Segmentation Model

We now describe a new model for melodic segmentation which identifies segmentation boundaries as perceived changes in melodic density. We will designate this model as Melodic Density Segmentation Model (MDSM). While the LBDM measures the accumulated boundary strength and identifies local maxima, in contrast the MDSM calculates the accumulated melodic cohesion between pitch intervals, and then identifies local minima (i.e. points of low melodic density) as local boundaries. This new segmentation method also incorporates a short-term memory window and models the effects of recency with an attenuation function.

Before a formal description of the model is presented, some of its characteristics and underlying assumptions must be explained.

4.2.1 Interval Order and Recency

It is conjectured that pitch intervals may be formed (and perceived) between all notes occurring over an interval of time (memory window) and not just between pairs of consecutive notes. In Table 4.1 an example of a short sequence of four midi notes is depicted together with the pitch distances (in semitones) between all pairs of events. The order of an interval determines the distance between the present and previous event considered. Thus, an interval of order k with respect to a given event e_i is denoted by (e_{i-k}, e_i). For example, from table 4.1 intervals (e_{i-1}, e_i) and (e_{i-2}, e_{i-1}) have order 1, intervals (e_{i-2}, e_i) and (e_{i-3}, e_{i-1}) have order 2, etc...

Recency effects apply in two different ways. The higher the order of an interval, the greater the temporal separation between the events, and therefore the weaker the perceived link between the two. On the other hand, more recently formed intervals have a stronger contribution to the melodic cohesion of the sequence than earlier formed ones. The recency of an interval with respect to an event e_i is given by the time that separates e_i and the latest event of the two that constitute the interval. These two factors are

Table 4.1: *Order* and *recency* of pitch intervals for a sequence of events. Intervals are in semitones.

e_{i-3}	e_{i-2}	e_{i-1}	e_i	*event*
53	50	50	48	*pitch*
				$order(n)$
	3	0	2	1
		3	2	2
			5	3
...	2	1	0	$recency(m)$

combined to determine the overall contribution of each interval at any given moment in time. In Table 4.1, recency is numerically represented in the bottom row. Increasing values of recency express less recent intervals. We suppose here, for simplicity, that all events in the previous example have equidistant on-set times and equal duration. Then intervals (e_{i-2}, e_i) and (e_{i-2}, e_{i-1}) will have equivalent contribution, since the former is an interval of order 2 (meaning that events are separated by 2 duration units) but with recency 0, and the latter has order 1 but recency 1 (meaning that the interval is separated from the reference event e_i by 1 time duration unit).

4.2.2 Melodic Cohesion

The melodic cohesion of an interval is defined here to be proportional to the frequency of occurrence of that interval in the interval framework associated with the melody being analysed. The use of interval frequencies to rate pitch intervals follows from some of the ideas introduced by the General Pitch Interval Representation (GPIR) (Cambouropoulos, 1998). But, while the GPIR defines three categories to rate the interval frequencies, in the MDSM frequencies have been converted into seven ratings. Additionally we have attributed to the unison the highest rating (7) and the octave the lowest rating (1). The lower rating for the octave interval implies that the MDSM does not consider octave equivalence, that is, it does not rate an octave as having high melodic cohesion, as discussed in Chapter 2. Intervals greater than the octave are given zero rating so their contribution to the melodic cohesion of a sequence is neglected.

In Table 4.2 we depict the interval frequencies, and the adopted MDSM interval ratings for all intervals in a major-minor scale framework [1].

[1]The GPIR major-minor framework is, as proposed by Cambouropoulos (1998), a weighted blend of the interval frequencies in major and minor scales.

Table 4.2: Interval frequencies and ratings for a major/minor scale framework

Interval name	unis.	2m	2M	3m	3M	4P	4a	5P	6m	6M	7m	7M	oct.	> oct
Pitch distance	0	1	2	3	4	5	6	7	8	9	10	11	12	>12
GPIR freq	–	0.32	0.65	0.57	0.43	0.76	0.19	0.76	0.43	0.57	0.65	0.32	–	–
MDSM rating	7	3	6	5	4	7	1	7	4	5	6	3	1	0

A short-term memory window determines the span of recent events that can form intervals. The width (duration in real time) of this window is fixed. The tempo of the piece will determine the number of recent events that can be recalled (within the window) and contribute to the melodic cohesion of the sequence.

4.2.3 Melodic Density

We can now formalise the notion of melodic density (MD) as the weighted sum of the contributions of all intervals occurring over a period of time determined by a memory window M. So given a sequence of N events $(e_1, ..., e_N)$ representing a melodic sequence the melodic density d_i at event i, is defined as:

$$d_i = \sum_{m=0}^{t_i - t_{i-m} < M} \sum_{n=1}^{t_i - t_{i-m-n} < M} f(r(e_{i-m}, e_{i-m-n})) \cdot a_i(m, n) \tag{4.1}$$

where $f(r) \in [0, 1]$ is a function that returns the frequency rating of an interval based on a given interval framework, and $r(e_j, e_k) \in 0, 1, ...12$, denotes a pitch interval in semitones, given by $r(j, k) = |p_j - p_k|$ where p_j, p_k denote the MIDI pitches of events e_j and e_k, and

$$a_i(m, n) = (1 - \frac{t_i - t_{i-m-n}}{M})^2 \tag{4.2}$$

is an attenuation function, where t_i denotes the onset time of event e_i, and M is the duration of the memory window (in seconds).

It is worth noting that the Gestalt-based principle of proximity is encapsulated in the attenuation function, with respect to time. The function will return values closer to 1 for recent and low-order intervals, and values closer to 0 for remote and high-order intervals.

Boundaries, in the MDSM, are indicated by local minima in the melodic density profile obtained from Equation 4.1. These minima correspond to points of low accumu-

lated melodic cohesion and, based on our assumptions, identify local boundaries in the melodic sequences.

4.3 Experiments and Results

To assess the behaviour of the model we used both the LBDM and the MDSM on a set of melody examples. For each of the examples we also obtained a pattern boundary profile, which indicates the location of recurrent patterns within the melodic sequence (see Cambouropoulos (1998) for details).

The interval frequency ratings given by function f were obtained from the combined major-minor framework shown in Table 4.2. The pieces were all assigned a tempo of 90 crochets/min and the memory window M was set to 4 seconds. The choice of this memory window size is based on reports of experiments on auditory recall (Eysenck and Keane, 1995; Snyder, 2000), that indicate average short-term memory retention times of 3-5 seconds.

Table 4.3 summarises the boundary counts for each melody, including pattern boundaries and the segment boundaries generated by both the LDBM and the MDSM [2]. A threshold of 70% was adopted to select only the most prominent peaks from the boundary profiles. A boundary is then marked correct if a pattern boundary exists within the distance of ± 1 event. Both these evaluation criteria were adopted for the sake of consistency, as they had previously been adopted by Cambouropoulos (1998) in his experiments.

From the analysis of Table 4.3 it can be observed that the LBDM generated 55 boundaries against only 50 by the MDSM. Both models correctly identify approximately the same number of pattern boundaries, but the MDSM is more selective, generating only 7 (14%) excessive boundaries , against the 16 (29%) of the LBDM. In the melodies where excessive boundaries were found, the MDSM always registers a lower count. However it must be noted that melody number 6 (theme of Mozart's Symphony in Gm), alone, is responsible for the majority of the excessive boundaries generated by the LBDM.

Since the distributions of the segment boundaries are non-normal, statistical parameters like mean and standard deviation, cannot be used to describe these sequences of boundaries and compare the performance of both models. The F-measure (Van Rijsbergen, 1979) can be used to rate the quality of a given alignment with respect to a reference,

[2] This table reflects two corrections in the boundary counts, which were detected posterior to its publication (Ferrand et al., 2003a). However these corrections have not changed the overall trend of the results

Table 4.3: Segmentation results obtained for 7 melodies, showing the total no. of pattern boundaries (PB), and for both the LBDM and MDSM: total no. of pattern boundaries found (fnd), no. of pattern boundaries not found ($notfnd$) and no. expurious boundaries found (ex)

		LBDM			MDSM		
Melody	PB	fnd	$notfnd$	ex	fnd	$notfnd$	ex
1. Lightly Row	5	5	0	0	5	0	0
2. Frère Jacques	7	3	4	0	5	2	0
3. Twinkle-twinkle	5	5	0	2	4	1	1
4. Yankee Doodle	5	3	2	3	5	0	2
5. L'Homme Armé	9	8	1	0	9	0	0
6. Mozart Symphony Gm	6	6	0	11	6	0	4
7. Beethoven 9th	9	9	0	0	9	0	0
Total	46	39	7	16	43	3	7

and it combines the notions of $Precision(P)$ and $Recall(R)$ in a single efficiency measure given here by their weighted harmonic mean:

$$F = 2 \times \frac{P \times R}{P + R} \tag{4.3}$$

where

$$P = \frac{PB_{fnd}}{PB_{fnd} + PB_{notfnd}}, R = \frac{PB_{fnd}}{PB_{fnd} + PB_{excessfnd}} \tag{4.4}$$

Precision and Recall have often been used to measure the efficiency of information retrieval methods (Manning and Schüttze, 1999), and the F-measure has been previously adopted to evaluate the performance of other melodic segmentation algorithms (Bod, 2001; Thom et al., 2002)

One of the characteristics of the F-measure is that it is rather unforgiving to errors and does not take into consideration near-alignments or approximate matches. Correct and incorrect alignments are obtained based on prior interpretation of the output of the segmentation algorithm. But if one wanted to consider boundaries of varying strengths or even compare a set of boundaries predicted by an algorithm with a set of possible reference segmentations, we would need some probabilistic notion of precision and recall. Thom et al. (2002) highlight the fact that, with the F-measure, two adjacent correct alignments have the same contribution as two correct alignments in distant locations and suggested the development of a probabilistic model that explicitly considers both seg-

Table 4.4: F-measure for the LBDM and MDSM

Model	*Precision*	*Recall*	F-measure
LBDM	0.85	0.71	0.77
MDSM	0.93	0.86	0.90

ment positions and boundaries (Spevak et al., 2002). For a detailed discussion of some of these alternative measures to precision and recall, the reader is referred to Raghavan et al. (1989).

In Table 4.4 we can see that although the MDSM only has a slightly higher *Precision*, it has a significantly higher *Recall* resulting in a higher value of F.

To further illustrate these results, in Figures 4.1 and 4.2, the boundary profiles of both models are shown together with the corresponding scores . For ease of comparison, the Melodic Density profile has been inverted [3] and both profiles normalised in the range 0-100%.

In the melody *Frère Jacques* we observe that some of the boundaries generated by the LBDM were eliminated due to the 70% selection threshold, although smaller peaks can be found in the vicinity of the pattern boundaries that were missed (intervals 4, 20 and 25 in Figure 4.1(a)). But, generally, an adjustment of the selection threshold to considerably lower values, will result in a increase of the number of peaks that are extracted, and consequently in an increase of the number of spurious boundaries. On the other hand, we would expect that an increase of the selection threshold would increase the selectivity of the model. But in the example of Figure 4.2 we can observe that this is not always the case. Most of the peaks in the LBDM profile have values over 80% and often in the 100% mark, thus making the elimination of the excessive boundaries difficult to achieve only by adjusting the selection threshold.

The example of Figure 4.2 also illustrates how some of the boundaries "filtered" by the MDSM correspond to weaker boundary locations and are not coincident with pattern boundaries.

[3] We recall that the MDSM segment boundaries are originally obtained from the local minima of the profiles

4.4 Discussion

The boundary selectivity reported on the MDSM, results partly from the propagation of interval information over a time window producing a "smoothing" effect which diminishes the contribution of smaller discontinuities of the melodic surface. In some cases, due to this effect, boundaries can be shifted forward or prolonged due to a slower decay of the melodic density function. This is visible in Figure 4.1 where the boundary peak after the third measure is followed by a significantly slow decay of the MDSM values (specially when compared with the sharp drop in the LDBM profile), until it meets the following peak. A similar effect can be seen in Figure 4.2(b) between the two boundaries occurring in measures 12 and 13, where boundary strength values never drop below the 70% threshold. This effect may have an impact on the accuracy of the boundary locations, particularly if matched against pattern boundaries without the adopted ± 1 event distance tolerance.

The choice of the attenuation function (a decaying polynomial), is the result of comparative preliminary experiments with the algorithm, where several attenuation functions were examined. However, it must be said, the differences were not conclusive. It seems intuitive that, in general, less recent notes have a smaller contribution to the melodic cohesion of a sequence, than more recent ones. However, to the best of our knowledge, there is no theoretical or experimental evidence to support the choice of a specific memory decaying function.

Although tempo was kept constant in this study, the MDSM is robust to small changes in tempo. This is mainly due to the discrete nature of the events, combined with a memory window of fixed size. For example, with a tempo of 60 crochets/min, a memory window of 5 seconds would include 5 crotchets (or the equivalent in duration). But it would be necessary to increase the tempo to 72 crochets/min (+20%) to include an additional crotchet in the calculations. Few studies have addressed the effects of changes in tempo in music memorization (Handel, 1993). Although the MDSM was designed to account for changes in tempo, a systematic evaluation of these effects has not yet been carried out. For such analysis we would require that listeners be tested on the effects of tempo changes to provide data to be compared with the model.

As mentioned previously, interval ratings were derived from the combined statistics of interval counts from major and minor scales. Since one of the motivations of this work is to devise a model that can segment melodies without any domain specific knowledge, we propose that these frequencies may be acquired from a corpus of music that is repre-

sentative of the melodies being analysed. This idea is supported by several studies, some of which were carried out outside the western musical culture, that report, for example, the prevalence of small melodic intervals in melodic lines (Bregman, 1990; Krumhansl, 1990; von Hippel, 2000). If indeed the melodic preferences of a particular musical culture are reflected in the musical material, it seems reasonable to reverse this process, by using implicit intervalic information to interpret the musical material.

To illustrate this idea, and subsequent to experiments with the MDSM, we looked at the frequency counts of melodic intervals in larger corpus of melodies. In Figure 4.3(a) pitch interval counts are shown for a subset of the Essen Folk-song Database (Schaffrath, 1994). When we compare these interval frequencies with those of the GPIR major/minor framework (see Figure 4.3(b)) we observe that there is a correspondence between the two, although in the Essen Folk-songs larger intervals tend to have significantly lower relative counts. Clearly, the GPIR interval frequencies are not exclusively related to melodic properties. They also express the harmonic properties of the major/minor tonal idiom, so it is not unexpected that intervals such as perfect fourth and fifth have such higher counts. The Essen Folk-song interval frequencies seem to be more closely (inversely) related to pitch distance, reinforcing the idea that Gestalt proximity rules in the pitch domain can emerge from the melodic material, and this could explain in part why they are prevalent in melodic perception. Furthermore it is suggested that interval frequencies could be learned from a musical corpus, thus avoiding the need for them to be introduced *a priori* in the model. Arguably this could make the model more adaptable as it is acceptable that different musical idioms might carry different intervalic distributions and implicitly different notions of melodic cohesion.

The MDSM follows the assumption that there is a relationship between melodic cohesion and the frequency of occurrence of pitch intervals found in the music material under analysis. In the example of Figure 4.3(a) it can be observed that there are significant differences between the frequency of occurrence of descending and ascending intervals (e.g. major seconds, perfect fourths). This could suggest that maybe direction of the intervals should be also accounted for by the model. Intuitively is seems difficult to argue that a descending sequence of major seconds has higher melodic cohesion that the equivalent ascending sequence. To the best of our knowledge no studies have focused on this particular issue. Perhaps these discrepancies should not be looked at in isolation and, as some research has suggested, more context is necessary (Krumhansl, 1990; von Hippel, 2000) to understand the underlying constraints between intervals that are used in melodic construction.

4.5 Summary

We presented the MDSM, a memory-based melodic segmentation algorithm based on the concept of melodic density. We compared this algorithm with the LBDM, for a set of melody examples. It was shown that, for these examples, the MDSM has higher selectivity than the LBDM, generating fewer total boundaries but preserving most boundaries that coincide with pattern boundaries. This supports the idea that the MDSM may be used successfully as a pre-processing segmentation method for pattern finding algorithms, providing in some cases additional reduction of the search space without the cost of eliminating many candidate pattern boundaries.

With the MDSM we proposed a novel approach to the problem of melodic segmentation, more data driven and where most parameters of the model are related to perceptual phenomena.

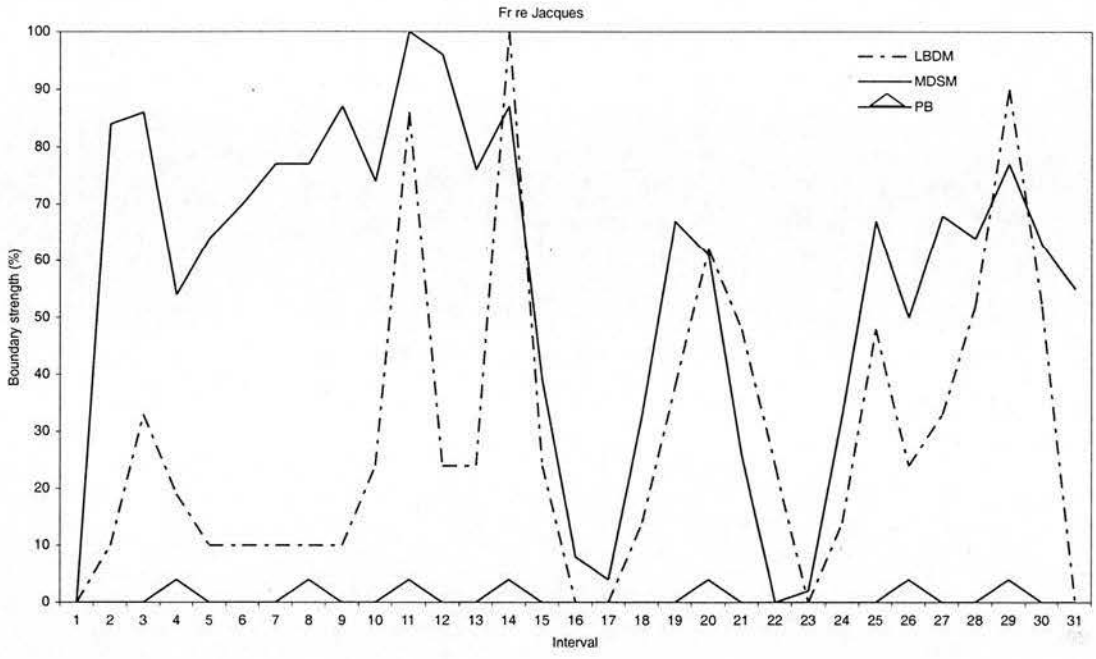
One particular contribution of this new approach lies in the way it incorporates pitch and time. Unlike the LBDM where time distances are calculated independently and then added to the pitch differences to determine the overall boundary profile, the MDSM concentrates on pitch differences and uses time to weight these differences in terms of continuity and temporal recency.

The emphasis on a time-based instead of an event-based representation brings the model closer to a real listening experience. Further experiments would be needed to fully explore some of these aspects of the model, such as varying the tempo of the melodies or the length of the memory window. But this would demand the realisation of listening experiments to produce data suitable for comparison with the corresponding model's segment predictions.

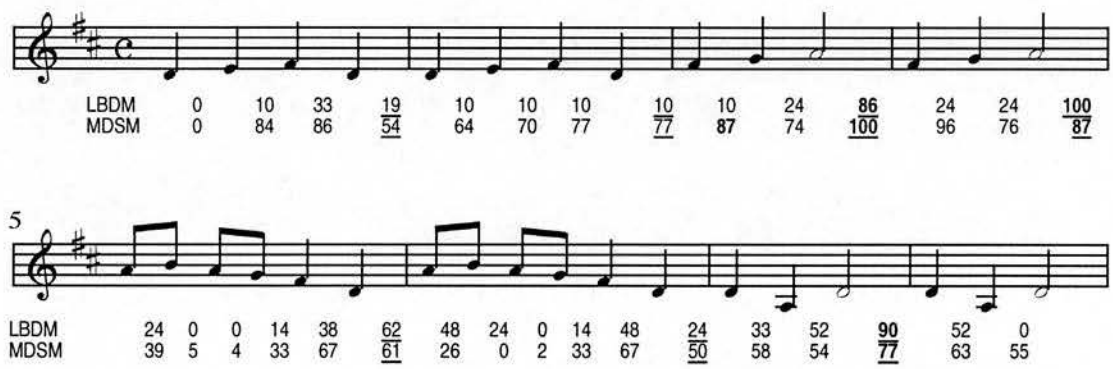
Pattern boundaries were used as a frame of reference to compare the performance of both LBDM and MDSM. Evidence shows that repeating patterns in a melody are often preceded or followed by large pitch intervals, changes in rhythmic patterns or rests. This may explain why so many pattern boundaries are coincident with the local segment boundaries generated by both models, which in itself don't contain any ability to detect parallelism or compare patterns that recur in distant sections in the same piece. Despite this existing overlap there is room for improvement by incorporating some form of melodic similarity in the model. On the other hand parallelism alone cannot account for all segmentation boundaries (Cambouropoulos, 2003), so perhaps a combination or an interaction between melodic similarity and local boundaries may prove more effective.

With the aim to continue the work on melodic segmentation, some implications and

new directions stem from this preliminary experiments with the MDSM. Firstly, segmentation may not be achievable in one pass and some form of memory model must be devised to store information that would correspond to learned information from a preliminary audition of the piece. The importance of learned musical information was here suggested by the use of statistical information from a corpus of music, to rate different pitch intervals. This emphasis on memory and learning from the musical data, is the main motivation for the development of a new probabilistic model for melodic segmentation, which is the subject of study of following chapters. Also, it seems important to address and discuss the interaction between Gestalt-based grouping rules, often associated with intrinsic human perceptual faculties, and parallelism, more associated with learning. Finally, if while modelling melodic segmentation we want to make claims about listening behaviour, it is paramount to collect listening data from real listeners. A listening study on melodic segmentation has been carried out and its results presented in the next chapter. We recall that the MDSM makes claims about the influence of tempo in perception so it would be desirable to address real-time issues in any further developments. Due to limited time and resources, real-time factors such as tempo, have not be addressed in this listening study and consequently have not been accounted for in further experiments included in this dissertation.

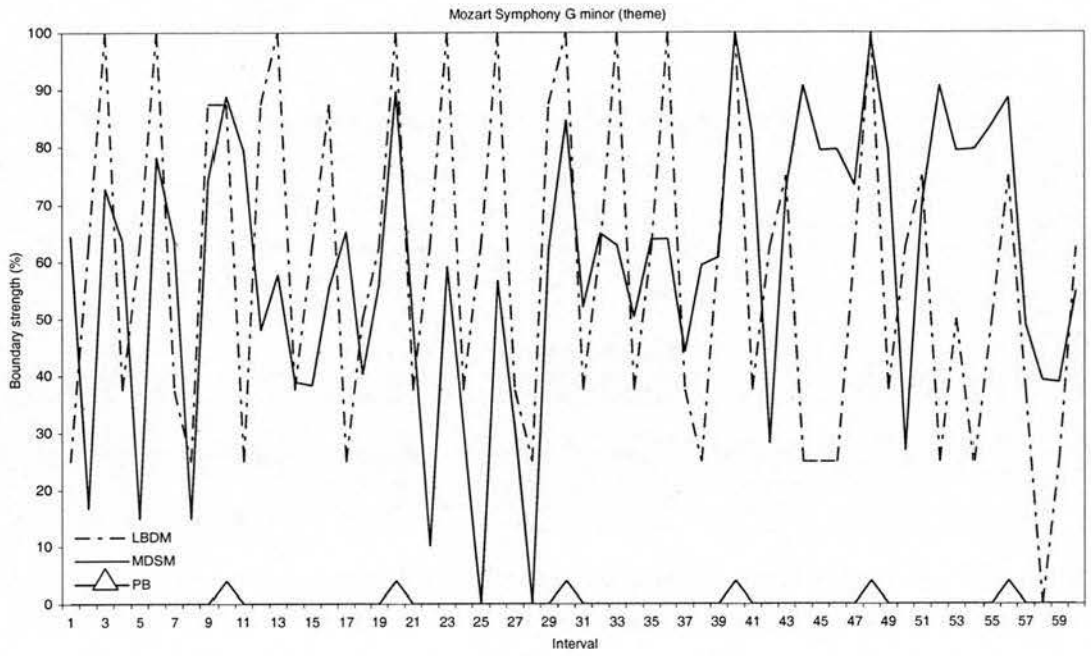


(a)

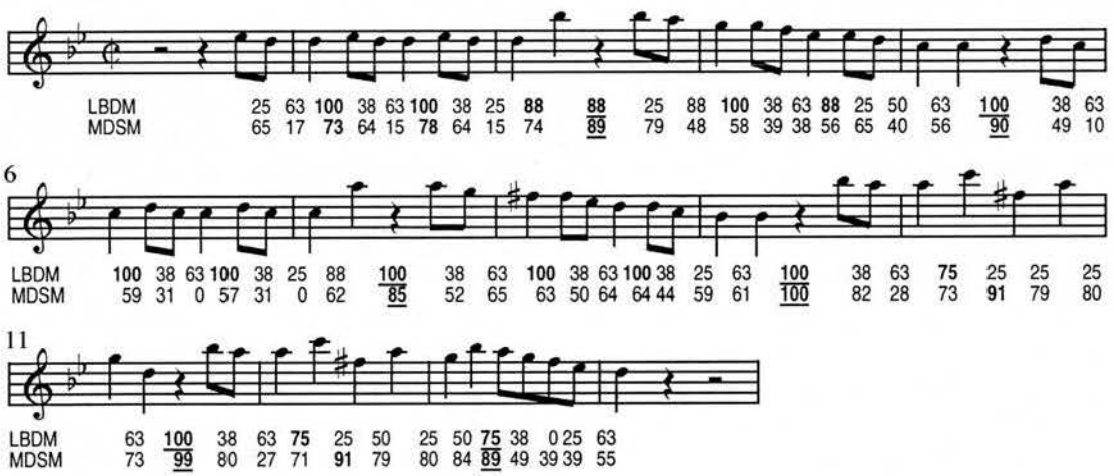


(b)

Figure 4.1: Melody number 2 : *Frère Jacques*. a) Boundary profiles for LBDM (dotted line) and MDSM (solid line). Pattern boundaries are indicated by arrows at the bottom of the chart. b) Score and superimposed boundary strengths. Selected boundaries are shown in boldface and values corresponding to pattern boundary locations are underlined

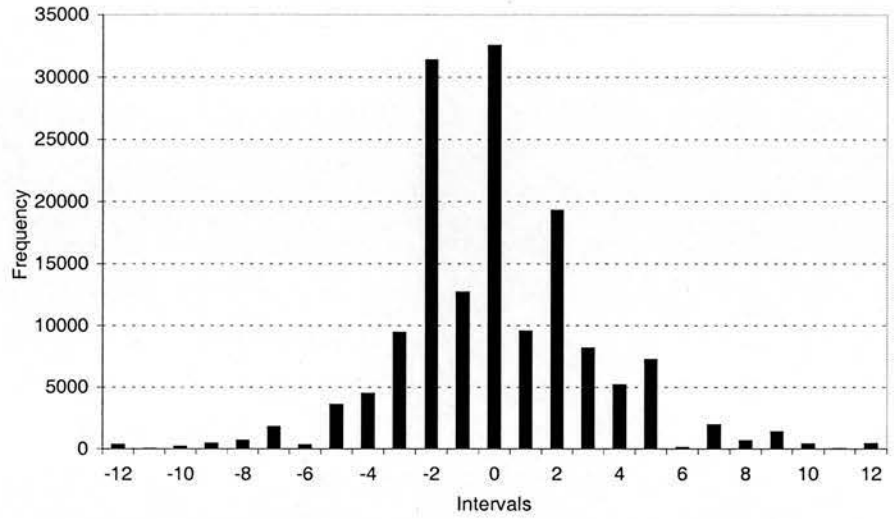


(a)

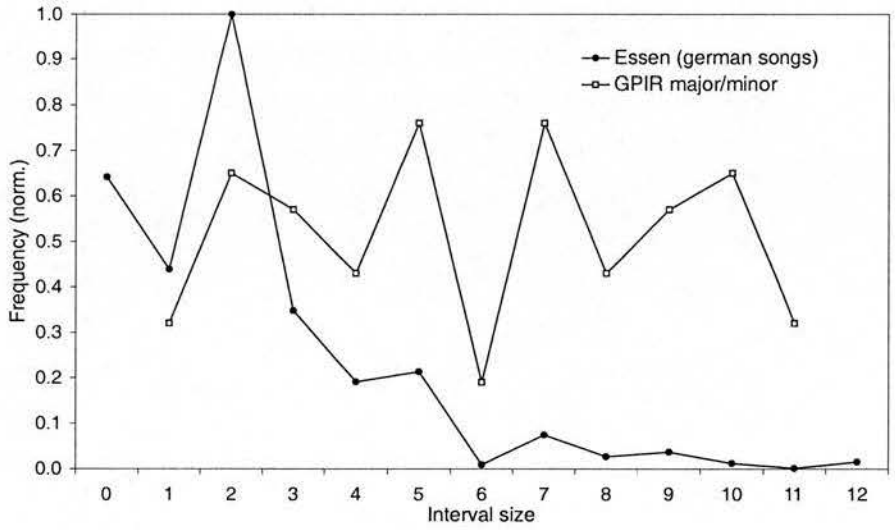


(b)

Figure 4.2: Melody no. 6: theme of Mozart's Symphony in Gm. a) Boundary profiles for LBDM (dotted line) and MDSM (solid line). Pattern boundaries are indicated by arrows at the bottom of the chart. b) Score and superimposed boundary strengths. Selected boundaries are shown in boldface and values corresponding to pattern boundary locations are underlined



(a)



(b)

Figure 4.3: Pitch interval frequencies (intervals denoted in semitones.) a) Interval distribution for the subset of German Folk-songs from the Essen Folksong Database. b) GPIR major/minor framework *vs* German Folk-songs interval frequencies

Chapter 5

An Empirical Study of Melodic Segmentation

When modeling listening behaviour, researchers are often faced with the need for real listening data to compare with the results of models or simulations. With this in mind, a study of melodic segmentation was carried out with several subjects on a few melodic examples. The aim of this study was to collect segmentation data for a corpus of melodies, from a realistic listening experience and thus to provide comparison data for our computational model of melodic segmentation.

5.1 Preliminary Segmentation Studies with Listeners

While a listening study was still being prepared for the purposes of the present research, a collaboration took place with the University of Bologna to organise some listening experiments in the University of Edinburgh. These experiments were part of a project carried out at the University of Bologna, for the study of *macroform* in post-tonal music. The term *macroform* is defined by the authors as the global form of a piece, that is, the division of the piece into its largest parts with reference to its overall structure (Addessi and Caterina, 2002).

The objectives of these experiments included the study of the relationship between the large-scale segmentation (*macroform*) perceived by subjects while listening to a piece in real-time, and the segmentation carried out by analysis of the piece by a group of musical experts. The study also aimed to investigate the correlation between the listeners' perception of tension/relaxation and the previously identified segment boundaries.

Not all of the aims of that study were relevant to the scope of the present research, but because it included a real-time segmentation experiment, this was an opportunity to become familiar with the organisation of a listening experiment of this kind. From

this collaboration resulted a reflection on several practical aspects of an experimental procedure with listeners, and this was paramount in the preparation of the listening study that would follow.

One of the aspects that can only be evaluated by experience is, for example, the length of the sessions and the consequent fatigue caused to the listeners. The duration of a listening session must be adjusted according to the complexity of the tasks and the nature of the musical repertoire. Because it is often so difficult to recruit participants for this type of experiment, it may be tempting to make the best out of the available listeners. However to ensure the quality of the data, the duration of the experiment must be carefully planned. Unfortunately time is often the main limitation on the amount of data that can be collected in a single experiment.

Providing the listeners with clear objectives and instructions for the experiment is another key issue. In the experiments carried out with the University of Bologna in Edinburgh, this became more evident as it was necessary to translate some of the materials (originally in Italian), such as the instruction sheets and questionnaires. In some cases it became necessary to adapt some of the terminology used.

The EPM software[1] used to play the pieces and collect the data from the listeners had all the menu captions and messages in Italian, making it less clear for the British listeners to manipulate without the aid of very precise instructions. Overall, due to the simplicity of the interface this did not interfere significantly with the experiment, but it was found that listeners were more prone to errors, because they were very dependent on the instruction sheets provided. For example, after each session, listeners had to access the menus of the application in order to manually save the results of the session, including typing the name of the file. On a few occasions results were saved with incorrect names, and data was accidentally overridden or lost. These occurrences showed that to guarantee the integrity of the data, it is advisable to remove technical operations from the participants and leave subjects to concentrate on the experiment itself.

The handling of the data also revealed a few points of concern. Typically, experiments of this kind require a large number of participants and whenever possible, it is useful to gather several participants in one single session. Open access computer labs can often provide the necessary space and number of machines required, but don't always have top of the range hardware. Multimedia resources such as audio files can be large in size, making them difficult to store, and slow to process on slower machines. It is important to ensure that the software can run smoothly, to avoid interruptions while playing

[1] EPM (Experiments on the Perception of Music) was developed in the University of Padua

Group	No. subjects	Age mix Mean(SD)	Gender mix F/M
Musicians	24	24.46 (3.30)	10/14
Non-musicians	24	24.29 (3.92)	14/10
Totals	48	24.38 (3.58)	24/24

Table 5.1: Participants' age and gender mix

sound files or delays in storing results because data is being logged and indexed in real-time. The format of the data logs should also be carefully planned, preferably stored in a format that is easy to import a database, where analysis of the data can be carried out promptly. Non-standard output data formats may add tedious hours of manual editing of the log files.

5.2 Description of the Study

5.2.1 The Participants

A total of 48 participants took part in this study. Participants were all postgraduate or final year undergraduate students, of mixed gender and with an average age of 24.38 years old. Participants were recruited and selected according to their musical training to constitute two equally numbered groups of 24 subjects with musical training and 24 subjects with no musical training. It was considered that subjects with musical training were all those with 8 or more years of formal musical studies and/or proficiency in a musical instrument. For ease of reference we will designate subjects with and without musical training as "musicians" and "non-musicians", respectively. The participants' age and gender mix is summarised in Table 5.1. A more detailed profile of the participants in this study can be found in Appendix C. All participants were given a book voucher as a remuneration for their participation.

5.2.2 Materials

For this listening study, five melodies were prepared in the form of Midi files. The pieces were Debussy's Syrinx (a solo piece for flute), two excerpts of melody parts of Mozart piano sonatas K284 and K333, and three German folk songs from the Essen Folk-song Collection (Schaffrath, 1994). Syrinx is an expressive (non-mechanical) performance se-

Designation	Source/description	Author	Duration
Syrinx	Flute solo, Integral	Debussy	2′ 14″
K284	Piano Sonata, Allegretto, Tema (bars 1-17)	Mozart	58″
K333	Piano Sonata, Allegro (bars 1-26)	Mozart	46″
E0547	Essen Folk-song Collection, Integral	anonymous	17″
F0927	Essen Folk-song Collection, Integral	anonymous	16″
Q0034	Essen Folk-song Collection, Integral	anonymous	20″

Table 5.2: Description of the melodies used in the listening study

quenced by Peter-Jan Van Dijk. The remaining pieces were typed in Sibelius music notation software, and then recorded as dead-pan Midi files.

A more detailed description of the listening set used can be seen in table 5.2, including all the durations of the pieces. For further reference, the scores of all the melodies used can be found in Appendix C.

This set of melodies was chosen according to carefully considered criteria, in particular, to include lengths ranging from short to long, different types of melodic, rhythmic and metrical patterning, possibly leading to different sets of segment structure.

It was not an aim of this study to carry out a systematic analysis of the effects of different pitch and rhythmic structures in segmentation. The main focus of the study was to acquire statistical information about perceived musical segments. The total number of songs used was mainly determined as a result of balancing a few of these different characteristics in the songs and keeping the duration of each listening session under an acceptable time.

Other criteria were also in mind when this melody set was chosen. The use of the Essen Folksong Database as a source of listening material, was motivated by the fact that many previous studies have focused on songs from this collection. These other studies may provide a valuable source of research data, to which our study may be compared. In particular, the three folk songs here chosen were included in a study on melodic segmentation (Thom et al., 2002) which we will refer to, later in this chapter. Similarly, the Mozart piano sonatas have also been used in another listening study relating segmentation to expressive timing (Cambouropoulos, 2001a).

A software application called Music Puncher was developed to play the melodies and collect the segmentation data from the listeners. The application was designed to control a whole listening session, providing the listeners with all the necessary guidance and instructions. All technical operations are kept away from the user and data logging

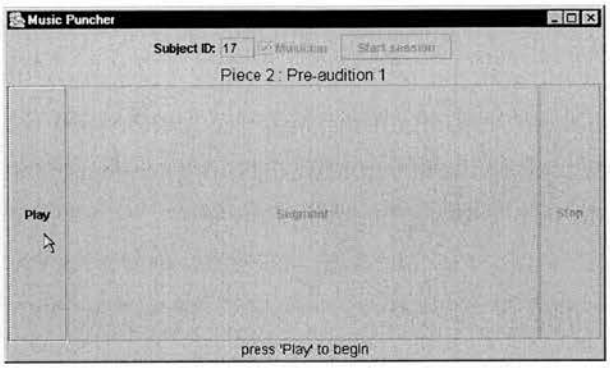


Figure 5.1: Screen-shot of the Music Puncher interface during a session

and storing are done automatically.

Music Puncher allows a subject to listen to a piece of music and to specify segmentation points by clicking the mouse button while the piece is being played. Segmentation points are recorded with a time stamp which specifies the number of elapsed milliseconds from the start of the piece. The interface was designed for maximum simplicity, both in terms of visual feedback (e.g. instructions given to listener) and in terms of interaction (see screen shot in Figure 5.1). For example it allows subjects to listen and press the large 'Segment' button without having to look permanently at the computer screen. This is an important feature as it was found that some participants preferred to carry out the segmentation task while listening with their eyes closed.

Music Puncher allows limited customisation of a listening session. One can provide a list of midi files to be played, and specify the number of familiarisation auditions allowed for every piece, as well as the number of practice segmentation auditions (prior to the final recorded segmentation audition). The application was developed in Java and can run on a Java Virtual Machine, allowing its use on different platforms and operating systems.

5.2.3 Procedure

The listening sessions took place in a computer laboratory with several PC/Windows NT workstations. Several listening sessions were organised each with a number of participants ranging from 4 to 8.

Each participant was assigned an individual computer and given a numbered instruction sheet/questionnaire (see a copy of this sheet in Appendix C). The sheet con-

tained a brief explanation about the purpose of the study, instructions on how to use the software, and a description of the segmentation task to be carried out. The sheet also included an information section regarding age, occupation and musical experience, to be filled by the subjects before the start of the listening session.

The segmentation task was described as follows:

> "You will be listening to 6 different melodic pieces. Try to imagine that each melody is a short story line that you have to break down in several smaller episodes (or segments)".

Once familiarised with the application interface, subjects were given a pair of headphones and instructed on how to adjust the volume of the sound to obtain their own level of comfort. Finally the session was initiated by the operator providing the subject number (as on the instruction sheet) to the application. Subjects were then asked to follow all the instructions on the screen.

For each one of the six melodies, participants were given two familiarisation auditions, followed by one practice segmentation audition and then a final segmentation audition. During the segmentation auditions subjects were allowed to click the mouse button any number of times, to indicate the locations of segment boundaries. Only the final segmentation audition data was recorded.

After completing each melody, the application would pause and subjects were instructed to remove the headphones and rest for a few seconds. The session is later resumed by the subject by pressing the "Start" button activated on the screen.

Sessions lasted approximately 45 minutes.

The melodies were presented in a different order to every participant. This was done automatically by the software assigning a unique permutation of the melodies based on the subject number entered in the beginning of the session.

After each listening session an informal debriefing was done to record the subjects' impressions regarding the listening session, such as difficulties felt in the face of the given segmentation task, the familiarity with the melodies played, etc.

5.3 Results

Figure 5.2 shows a sample of the data logged during a listening session by one of the participants. The first line indicates the subject number (unique to every participant),

Melody info	name	Syrinx	K284	K333	E0547	F0927	Q0034
	nº events	311	202	209	39	45	57
	duration (s)	133.9	57.5	45.5	16.8	15.6	19.8
All subjects	mean	10.8	8.0	7.2	2.6	2.7	3.3
	stdev	7.9	4.7	4.7	1.9	2.0	1.8
	min/max	0/36	2/33	1/21	1/12	0/7	1/8
Non-musicians	mean	7.9	6.1	5.1	2.5	2.6	2.6
	stdev	4.3	2.2	3.3	2.3	2.2	1.7
	min/max	2/17	2/10	1/16	1/12	0/7	1/8
Musicians	mean	13.6	9.9	9.3	2.7	2.8	4.0
	stdev	9.5	5.8	5.1	1.5	1.8	1.7
	min/max	0/36	4/33	2/21	1/7	1/7	1/8
Signif. of diff.	t(df)	t(32)=2.66	t(30)=3.04	t(39)=3.39	t(39)=0.23	t(45)=0.36	t(46)=2.90
	p	p<0.05	p<0.01	p<0.01	**p=0.82**	**p=0.73**	p<0.01

Table 5.3: Boundary segment counts: Top row of the table indicates number of Midi events and duration of each melody. Middle rows show average, standard deviation and minimum/maximum number of segments for all subjects, and for non-musician and musicians groups. Bottom row depicts the significance of the difference between the two groups

followed by a letter indicating the subject's musical training (here 'N' denotes a Non-musician). The data lines that follow contain the segmentation data, resulting from the mouse clicks. Every segment boundary is logged with the name of the corresponding piece followed by a timestamp (in milliseconds) indicating the elapsed time since the start of the piece.

28, N	...	...	...
e0356, 8161	syrinx, 49902	k284, 7681	k333, 5288
q0034, 10415	syrinx, 56251	k284, 14401	k333, 9614
q0034, 15843	syrinx, 67587	k284, 21000	k333, 16414
f0927, 8622	syrinx, 75639	k284, 27870	k333, 21782
syrinx, 13369	syrinx, 86464	k284, 35251	k333, 25697
syrinx, 36082	syrinx, 93545	k284, 42672	k333, 29533
syrinx, 45966	syrinx, 115586	k284, 50252	k333, 45976
...	...	...	

Figure 5.2: Example of data logged by one participant in a listening session

In Table 5.3, we can observe that there is a significant variability in the number of segment boundaries recorded across all participants, for each one of the pieces. For example, the total number of boundaries indicated by listeners range from 0 to 36 for Syrinx and from 2 to 33 for K284[2].

[2]We recall that subjects were not instructed to stay within any maximum or minimum number of boundaries. So participants could indeed indicate 0 boundaries, expressing in their own view, that the whole piece

Table 5.3 shows that on average Musicians indicate a higher number of segment boundaries than Non-musicians. This is particularly visible for the three longer melodies, Syrinx, K284 and K333. An independent T-test (assuming unequal variances) depicted in the bottom row of the table, indicates that this difference is significant for all melodies (at $p<0.05$ for Syrinx, and at $p<0.01$ for K284, K333 and Q0034), with the exception of E0547 and F0927, expectedly the two shorter pieces in the listening set. As the length of the melody increases, more data is available for comparison, so the probability increases that significant differences will be detected.

5.3.1 Segment Boundary Probability Density Estimation

Since segment boundaries are expressed by a time point relative to the start of a melody, it is possible to locate each boundary within the onset times of the midi events of that melody. Midi events are the only available frame of reference to look for any existing consensus regarding the placement of segment boundaries. This information could be used to generate a histogram of segment boundaries, using inter-onset intervals between Midi events as bins. However, the use of different sized bins seems inappropriate: for example, due to the response time of the listeners, bins corresponding to events of shorter duration could probably register low counts, failing to include some of segment boundaries to neighbouring bins. On the other hand, longer events could include boundary counts beyond relevance. Similar effects can occur near the boundaries of the bins. For example a boundary that is placed nearly at the onset of an event, would be counted only in one of the bins (before or after the onset), probably only due to a matter of milliseconds.

Although widely used as a density estimation method, histograms are known for their variable appearance, depending on a particular choice of origin and bin size. This is problematic since we are assuming that our data may be affected by non-negligible response times, with unknown direction. By unknown direction we mean that it is not possible to determine if, when clicking the mouse, listeners anticipated or responded late to the stimuli.

Therefore, we need a way of assessing the consensus of the listeners' segment boundaries without having to rely on a pre-defined time grid, or a set of initial likely boundary locations.

---

was to be taken as an indivisible segment

Kernel Density Estimation (KDE) (Silverman, 1986) provides a density estimation method suitable for the analysis of sets of time-indexed observations(Rieke et al., 1997; Toivianen and Snyder, 2003).

Let us consider that $t_k, k = 1, ..., n$ denote all the segment boundary times indicated by all the participants for a given melody. We use a Gaussian kernel density estimator (see Appendix B for a detailed explanation of KDE) to estimate the segment boundary probability density $p(t)$, given by

$$p(t) = \frac{1}{nh\sqrt{2\pi}} \sum_{k=1}^{n} e^{\frac{-(t-t_k)^2}{2h^2}} \tag{5.1}$$

The probability density $p(t)$ is thus a sum of Gaussian pulses (kernels) placed at each segment boundary time point. The choice of a Gaussian kernel follows the intuition that observed boundaries should be weighted as a function of their proximity to the reference stimuli.

The value of the bandwidth h has an effect on the curve of $p(t)$, introducing more or less smoothing of the curve, particularly in the regions where several contiguous boundaries are observed. The choice of a smoothing parameter is influenced by the purpose for which the density estimate is to be used (Silverman, 1986). In this context h could be interpreted as a maximum for the time response of the listeners, that is, the time difference between segment boundaries and the audio stimuli. Under the assumption that most boundaries are indicated in the vicinity of an event onset or offset, we determined the time difference between every recorded boundary and its closest midi onset/offset. A value of $h = 150ms$ was adopted, which is within the estimates and which seemed to provide a good trade-off between smoothing of the distribution and resolution. It is important to note that we are not making any claims about the significance of the adopted value of h, as being a perceptual constant of some kind. Although we believe that it may be related to the response times of the listeners, it is likely that it would vary depending on the experimental conditions and tasks involved. Ultimately, the aim of these estimates was to provide some guidance in the choice of a smoothing parameter, relating it to the listening data, and reducing the subjectivity of that choice.

In Figure 5.3, we show the segment boundary density curves for melody F0927 and in Figure 5.4 the same graphs are overlapped with the score of the melody. The plots of the density curves for all melodies user in the study can be found in Appendix C.

A comparison of the boundary density curves between groups shows that the profiles for Musicians correlated positively with those of Non-musicians. This correlation is

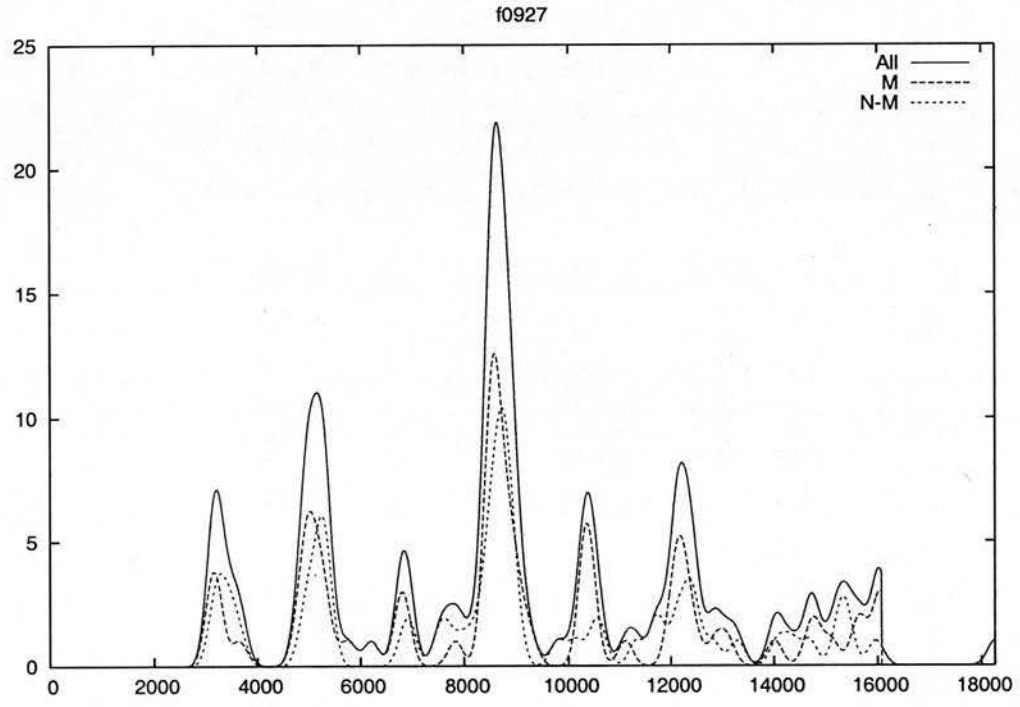


Figure 5.3: Segment probability density estimation for melody f0927. Solid line corresponds to the density estimation for all subjects, and the dashed and dotted lines correspond to the density estimation for Musicians and Non-musicians, respectively.

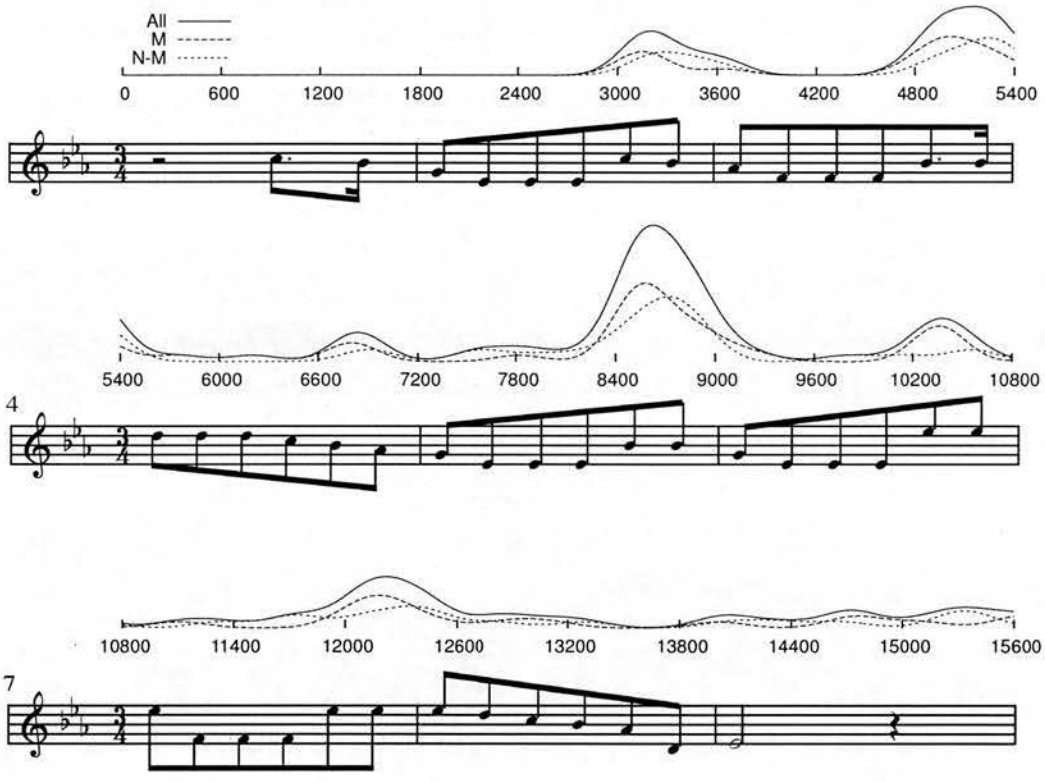


Figure 5.4: Segment probability density for melody f0927. The score is temporally aligned with the scale of the graph, depicting onset times (in milliseconds) for every quarter-note

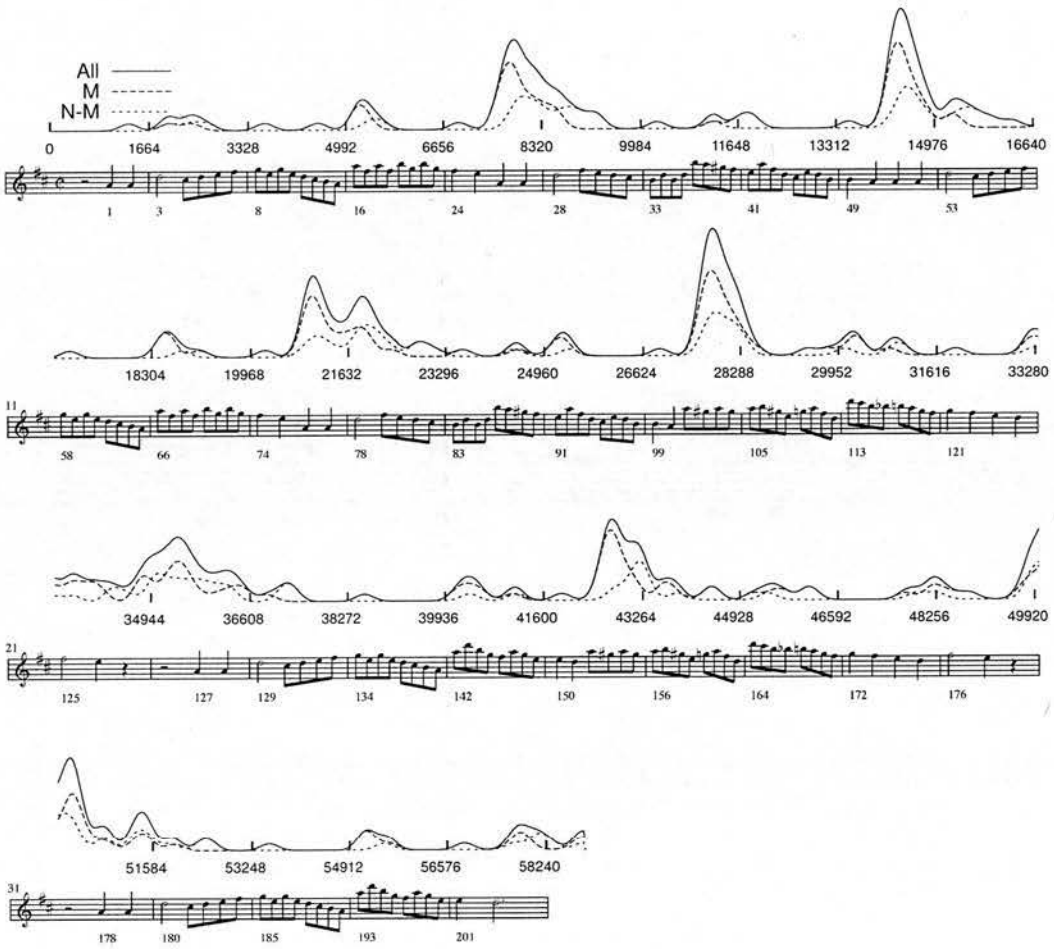


Figure 5.5: Segment probability density for melody K284. The score is temporally aligned with the scale of the graph, showing onset times (in milliseconds) for the first event of every bar.

significant (at $p<0.01$) for all pieces except for melody E0547 (see table 5.4).

5.4 Discussion

Overall, musicians seem to indicate a significantly larger number of segment boundaries, and thus smaller segments. So, are musically trained subjects more sensitive to certain musical features, producing as a result more elaborate segmentations? We know that musicians, as a result of their formal musical training, may have more elaborate and varied notions of what a segment might be in a musical context. But one could argue

melody	r	p
syrinx	0.709	0.000 (< 0.01)
k284	0.684	0.000 (< 0.01)
k333	0.680	0.000 (< 0.01)
e0547	**0.761**	**0.659**
f0927	0.854	0.000 (< 0.01)
q0034	0.915	0.000 (< 0.01)

Table 5.4: Correlations between segment boundary probability density profiles for musician and non-musician subjects

that such skills could work both ways. Some musicians might identify several sub-levels of segmentation in the melodies but choose to report only a few higher-level boundaries. However it is not an aim of this research to address this question. In this study, listeners were unable to express any kind of hierarchy associated with the segment boundaries they indicated. This may have led some of the listeners to decide a priori what level of segmentation they would convey, and in some cases they may have deliberately omitted lower level boundaries.

Although there may be differences in the way individual subjects perceived the different melodies, there is a significant positive correlation between musicians and non-musicians with the exception of one of the pieces (E0547). A close look at the segmentation of melody E0547 (see Figure 5.6) shows that even here the discrepancies are fairly localised. Non-musicians have indicated boundaries approximately after events 7 and 17, possibly to mark a sub-division of the longer phrases from bars 1-4 and 5-8, but the same sub-divisions were omitted by most musician participants. On the other hand mainly only musicians indicated a boundary after event 29, which coincides with the repetition of a motif in bars 9-10 and then 11-12.

For both musicians and non-musicians there is a visible spread of segmentation boundaries in the vicinity of the 'most voted' boundary locations. These could be attributed to anticipations or delays in the response to the melody. In some of the boundary density profiles (see for example Figure 5.5) the higher peaks have a higher value for musicians and the peaks are usually sharper. This may be due to the fact that musically trained subjects are more used to anticipating events within a piece of music and are therefore more accurate, when setting a boundary location with the mouse, in real time. These differences in response time or anticipation are difficult to estimate, but are visible in some of the density curves. For example, in Figure 5.3 we can observe a slight dis-

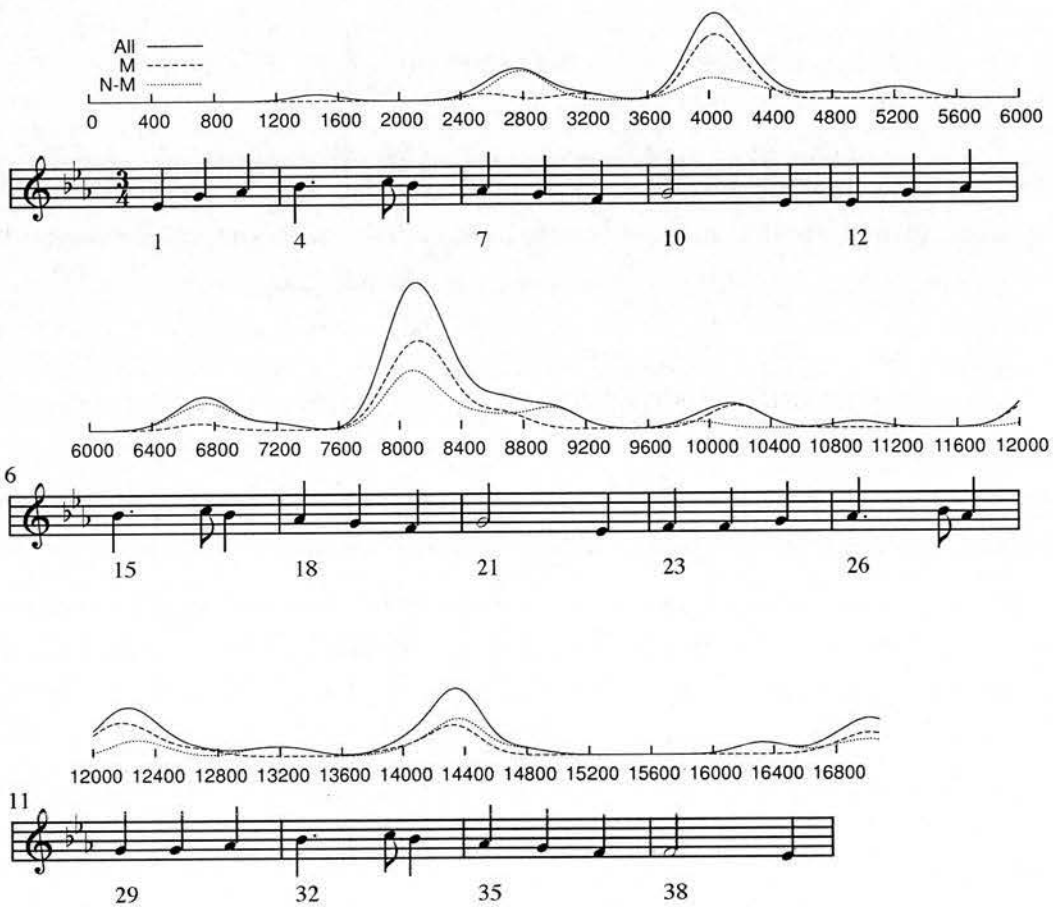


Figure 5.6: Segment probability density for melody E0547. The score is temporally aligned with the scale of the graph, showing onset times (in milliseconds) for the first event of every bar.

placement between the higher peaks. The peaks in the non-musician profiles are slightly shifted forward in accordance with the intuition that non-musicians are likely to have slower response times.

Are the boundaries marking the end of segments or the beginning of segments? This question is hard to answer: probably subjects had different strategies to indicate the boundary locations. To some extent the two strategies are equivalent in terms of the data that is obtained. Whatever the case may be, it is impossible to account for these individual options of all subjects and reflect them in the analysis. Also time responses may also vary between subjects. As a result, we would expect a residual distribution of boundary segments around the more consensual boundaries. This is confirmed by some of the density profiles where often higher peaks are prolonged slightly near the base. See for example in Figure 5.5 the peaks near events 26, 51 and 102, or in Figure 5.6 the peaks after events 20 and 29.

The familiarity of the subjects with the melodic examples used could not be determined until the end of the listening sessions, but this was not a criterion for the selection of the participants in the listening study. Listeners were asked informally, at the end of the sessions if they recognised some of the melodic samples and very few, even amongst the musically trained, had recognised some of the samples. The Mozart extracts were more often mentioned as being familiar, however it is not clear whether these subjects could in fact recall the particular movement of the sonata or were just reporting a familiarity with Mozart's style. Syrinx, was also somewhat surprisingly unfamiliar to most participants.

5.5 Summary

This chapter presented an empirical study on melodic segmentation carried out with two groups of listeners, with and without formal music training. Melodic segmentation is a widely subjective matter, and we recall that in this study, the notion of the segment was left open to interpretation. By using a large number of subjects we aimed to improve the estimation of a few common boundary locations for each one of the melodies provided as stimuli. Using probability density estimates it was shown that there are several prominent boundary locations, revealing a significant agreement between the two groups of listeners, regarding the segmentation of the melodies.

The data collected in this study, identifies a set of perceived prominent boundary locations for each one of the melodies in our corpus. This data will be used a reference

to test a new computational model of melodic segmentation which is presented in the next chapter.

Chapter 6

A Probabilistic Model of Melody Segmentation

This chapter presents in detail a computational model for melodic segmentation which is based on a probabilistic learning paradigm. We discuss the adopted representation for melodic information before we show how a mixed-memory Markov model is trained from the resulting melodic feature sequences. Finally we present our main hypotheses that are supported by information theoretic concepts, to make predictions about the location of segmentation boundaries, and we look at an example to illustrate how the model works.

6.1 Melodic Representation

Melodies can be seen as a temporal process where sound events unfold in time. In this work melody information is obtained from a MIDI source and then converted into an event-based symbolic representation, which approximates such a temporal process. A basic melodic event includes information about pitch, onset time and duration, and can be represented as follows:

Event A melodic event e is a triple $(Pitch, Onset, Dur)$ where:

- $Pitch(e) = Pitch$ represents the MIDI note number of the event
- $Onset(e) = Onset$ denotes the onset time of the event (in milliseconds).
- $Dur(e) = Dur$ represents the duration of the event (in milliseconds).

A melody can then be represented as a sequence of events $(e_1, e_2, ...e_N)$ which are numbered sequentially.

From this basic event information, two features are derived and associated with each event:

Pitch step (PS) expresses the interval distance in semitones from the previous to the present event:

$$PS(e_i) = Pitch(e_i) - Pitch(e_{i-1})$$

Duration ratio (DR) expresses the ratio between the two time durations associated with the present and previous event and is defined as:

$$DR(e_i) = IOI(e_i, e_{i+1})/IOI(e_{i-1}, e_i)$$

where $IOI(e_{i-1}, e_i)$ denotes inter-onset interval (IOI) between events e_{i-1} and e_i and is given by $IOI(e_{i-1}, e_i) = Onset(e_i) - Onset(e_{i-1})$

The choice of these particular two features is supported by empirical research overviewed in Chapter 2, which highlights the relative importance of certain melodic attributes in melodic recognition and similarity. Both chosen features PS and DR, represent melodic information in a relative manner, avoiding the use of absolute pitch values or absolute durations. This has the advantage of allowing some distinct melodic sequences to share the same representation whenever their events share identical intervallic relations or proportional duration, as we explain further on.

Pitch step can assume values larger than 12 semitones, implying that we have not considered octave equivalence but instead just a linear pitch scale expressing the distance between events. As we will see later in this chapter, for practical purposes, a limit of 2 octaves was imposed on the maximum size of a pitch interval. In other words, all intervals greater than 2 octaves will be rounded to that limit. It turns out that that the occurrence of such large intervals is rare, and in our corpus of music non-existent. The important point to retain is that an octave will be considered as an interval of 12 semitones, and similarly a major second (2 semitones) will be distinguished from a major ninth (14 semitones). The PS feature also implies that two melodic sequences will have the same representation as long as one is the exact chromatic transposition of the other.

Duration ratios represent intervallic information in the time domain. Unlike pitch intervals that are represented by an integer number, duration ratios are represented as a decimal quantity. These ratios can vary considerably particularly if we want to represent expressive (non-mechanical) performances of melodies, where the durations of MIDI events do not correspond exactly to the notated durations.

Rests are not addressed independently in our representation, since we do not considered rests as melodic events. Several models of music perception have treated rests as independent occurrences (Lerdahl and Jackendoff, 1983; Cambouropoulos, 1998). Instead we opted for defining DR, not in terms of the ratio between the durations of two events, but as the ratio of the inter-onset intervals (IOI) between events. This means that whenever a rest follows an event, its duration is implicitly added to the duration of that event.

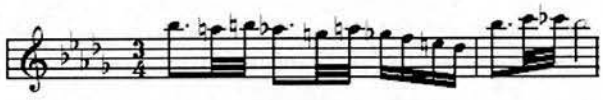
In Figure 6.1 we show the melodic representation for the first two bars of Debussy's Syrinx, together with features vectors for PS and DR. The DR values (in parenthesis) are also expressed in a logarithmic scale (rounded to the nearest integer). This converts the decimal ratios into a finite alphabet of integers more suitable to our model, as will be made clear in the next section.

The choice of a base 2 logarithm was intended to produce quantised ratios in the range $(-5, -4, ..., 0, ...5)$ for most melodies. This range corresponds to the intuition that listeners are unable to discriminate more than 5 ratios of change (not including equality). To the best of our knowledge, the ratings of sound duration ratios have not been studied empirically, but there is support (Miller, 1956) for the fact that humans can only perceive and distinguish a limited number of magnitudes.

It is acknowledged that the chosen melodic representation is a reduction of the auditory information conveyed to the listener. As such this choice of features in not by any means a conjecture about how music is perceived by the listener. We recall that the aim of this work is to compare results with a listening study in which only dead-pan MIDI files were played to the listeners, thus eliminating the influence of auditory elements such as timbre, loudness and other expressive elements. In face of the complex problem of music representation, we propose here a set of melodic features, encoding both pitch and time information, contained in the sources. The adopted melodic features are not intended to support the argument for a particular melodic representation, but rather to be a starting point in the research of the perception of melodic information.

6.2 The Memory Model

Our machine learning model is implemented using a mixed-memory markov model (see Section 3.2.3). The input to the model is a set of feature vectors containing sequences of symbols. In our case two feature vectors PS and DR are provided for each melody. A model is constructed for every feature vector by storing the counts of all existing m-



(a)

Id	Pitch	Onset	Dur	PS	DR	$\log_2$ DR
1	82	3998	1000	-	-	-
2	81	4998	142	-1	0.14	-3
3	83	5141	115	2	0.81	0
4	80	5256	987	-3	8.58	3
5	79	6256	167	-1	0.17	-3
6	81	6423	166	2	0.99	0
7	78	6589	273	-3	1.64	1
8	77	6862	322	-1	1.18	0
9	76	7184	388	-1	1.20	0
10	73	7577	468	-3	1.21	0
11	82	8048	1000	9	2.14	1
12	84	9048	148	2	0.15	-3
13	83	9197	120	-1	0.81	0
14	82	9318	3444	-1	28.7	5
..	..	...	...	.	.	.

(b)

Figure 6.1: Example of melodic representation a) The first two bars of Syrinx, b) Event information and feature vectors

id	pitch	onset	dur	id	pitch	onset	dur	id	pitch	onset	dur
1	63	0	125	21	62	4000	125	41	70	8000	125
2	62	125	125	22	60	4125	125	42	69	8125	125
3	62	250	250	23	60	4250	250	43	69	8250	250
4	63	500	125	24	62	4500	125	44	72	8500	250
5	62	625	125	25	60	4625	125	45	66	8750	250
6	62	750	250	26	60	4750	250	46	69	9000	250
7	63	1000	125	27	62	5000	125	47	67	9250	250
8	62	1125	125	28	60	5125	125	48	62	9500	500
9	62	1250	250	29	60	5250	250	49	70	10000	125
10	70	1500	500	30	69	5500	500	50	69	10125	125
11	70	2000	125	31	69	6000	125	51	69	10250	250
12	69	2125	125	32	67	6125	125	52	72	10500	250
13	67	2250	250	33	66	6250	250	53	66	10750	250
14	67	2500	125	34	66	6500	125	54	69	11000	250
15	65	2625	125	35	63	6625	125	55	67	11250	250
16	63	2750	250	36	62	6750	250	56	70	11500	250
17	63	3000	125	37	62	7000	125	57	69	11750	125
18	62	3125	125	38	60	7125	125	58	67	11875	125
19	60	3250	250	39	59	7250	250	59	65	12000	125
20	60	3500	500	40	59	7500	500	60	63	12125	125
								61	62	12250	250

Figure 6.2: Event list for the opening theme of Mozart's Symphony in Gm

separated pairs of symbols. Thus for each m, a separate table is created, holding pairwise dependencies (the contexts) of mth order. The maximum order of the model n determines the total number of symbol transition tables produced for each feature.

Transition tables have been implemented as simple arrays of numbers. Since we are only storing pairs of values, and each table is at most $k \times k$ (where k is the size of the alphabet), there is no need to use any space-efficient data storage structure.

An example of transition tables is shown in Figure 6.3, for the opening theme of Mozart's Symphony in Gm. In Figure 6.2 we depict the event list for this melody (see score in Figure 4.2).

Transition tables are used to estimate sequence probabilities but on its own they highlight some characteristics of the piece. For example, from the tables in Figure 6.3, it stems out the general predominance of sequences of descending pitch intervals. Note also that the recurring opening three-note pattern, is expressed in the first-order PS table by the high counts of pairs $(-1, 0)$ and $(-2, 0)$ and in the equivalent DR table by the pairs $(0, 1)$. In this example, DR assumes only one out of three distinct ratios, not surprising given the low rhythmic diversity of the melody.

The coefficients ϕ_m of the mixture model can be estimated using the Expectation-Maximisation iterative procedure referred to in the previous chapter. On every iteration

PS: m=1

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	
-6	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2
-5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
-2	0	1	0	0	3	3	6	0	0	1	0	0	0	0	0	0	0	0	14
-1	0	0	0	0	3	0	8	0	0	0	0	0	0	0	0	0	0	0	11
0	0	0	0	1	3	2	0	2	3	2	0	0	0	0	1	1	0	1	16
1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3
3	2	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	5
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1

DR: m=1

	-2	-1	0	1	
-2	0	0	5	0	5
-1	0	0	9	0	9
0	0	1	9	16	26
1	5	8	2	4	19

DR: m=2

	-2	-1	0	1	
-2	0	0	0	5	5
-1	0	0	1	8	9
0	1	9	9	6	25
1	4	0	15	0	19

Mixture coefficients for DR:
c[1]= 0.99999840
c[2]= 1.21237239E-6

PS: m=2

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	
-6	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
-5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
-2	0	0	0	0	2	3	3	0	3	0	0	0	0	0	1	1	0	0	13
-1	0	0	0	1	2	0	2	2	0	2	0	0	0	0	1	0	0	1	11
0	2	0	0	0	6	6	2	0	0	0	0	0	0	0	0	0	0	0	16
1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3
3	0	1	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	5
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
9	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1

Mixture coefficients for PS:
c[1]= 0.99999962
c[2]= 3.54766909E-6

Figure 6.3: Transition tables and mixture coefficients for PS and DR (1st and 2nd order) for the the opening theme of Mozart's Symphony in Gm. Each row indicates, for a given context symbol, the counts of all m-separated successor symbol occurrences. Total of occurrences in each row is given on the right margin of each table.

we compute:

$$\phi_m = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi_m \cdot a^m(w_i|w_{i-m})}{\sum_{l=1}^{n} \phi_l \cdot a^l(w_i|w_{i-l})} \tag{6.1}$$

The procedure iterates until the difference between all consecutive values of ϕ_m is lower than a pre-defined ϵ. In our implementation the termination condition was set to $\epsilon = 10^{-6}$. Also for practical reasons, an absolute maximum order $n = 10$, was imposed. As it will be shown in the next two chapters, the contribution of symbol dependencies of very high order will prove to be negligible, for all our melodic examples. The effective order of our mixed-memory model will be determined by the non-negligible mixture coefficient of higher order.

The mixture coefficients for our melody excerpt from Mozart's Symphony in Gm, are indicated in Figure 6.3 below the transition tables, and reveal that both memory models (for PS and DR) have an effective order equivalent to a bigram model.

As a result of the training of the memory model and using Equation 3.10, it is possible to determine the probabilities of a given feature sequence and make predictions about what feature symbols are likely to follow. This learning process is our model's equivalent of the familiarisation audition of a melody.

6.3 Segment Boundary Prediction

We argue that feature salience in a melodic sequence is strongly associated with the inter-opus distinctiveness of that feature. We conjecture that listeners respond to this salient features, and are likely to indicate segment boundaries at these locations. Feature salience is measured using an information theoretic approach and is associated to the changes in prediction entropy with respect to that feature.

Hypothesis Accentuated changes in prediction entropy over time may be associated with the perception of segment boundaries.

At any given point in a melody, a sudden decrease in entropy implies a reduction in ambiguity. From the listener's point of view, this decrease in ambiguity expresses the recollection of a previously memorised context and successor feature symbols.

Based on Shannon's notion of *source entropy* expressed in Equation 3.2, we can determine the entropy associated with any given context c as:

$$H_c(X) = -\sum_{\forall x} P(x|c) \log_2 P(x|c) \tag{6.2}$$

where x denotes all the possible successors of context c. Context c is a feature sequence of length $n-1$ for a model of $(n-1)$th order. Conditional probabilities $P(x|c)$ are obtained from Equation 3.10 and will reflect the statistics of the training set.

Note also that since the target melody and the training set are the same, no unseen events can appear as successors of a given context, therefore entropy values reflect all the available statistics. Given a context, the more equal the probability of its successors, the higher the entropy H_c will be. Conversely, entropy will decrease if the probability of a successor is very high (or very low), and the probability of the remaining alternative successors comparably low (high).

Typically, when the context coincides with the end of a frequent pattern, entropy is likely to be higher due to greater uncertainty about the successor of the progression. If, on the other hand, the context coincides with the beginning of a recurrent pattern, entropy is more likely to drop since the continuation of the pattern has been established with high probability successors. This alternation between high and low entropy, also reported by Witten et al. (1994), seems to occur in the transitions between recurring melodic phrases. In general, however, melodies are not so structured, so differences in entropy are expected to be more subtle.

Entropy vectors can be generated from Equation 6.2 by taking all the successive context sequences from each one of the feature vectors for the target melody.

As mentioned previously, we are interested only in the more accentuated entropy changes across the melody. More specifically, for every entropy vector, we consider only those points where entropy drops below the average entropy $\overline{H_c}$. This simple feature salience criterion can be given the function $S(i)$ which is defined as:

$$S(i) = H_{c_i}(X) - \overline{H_c} \tag{6.3}$$

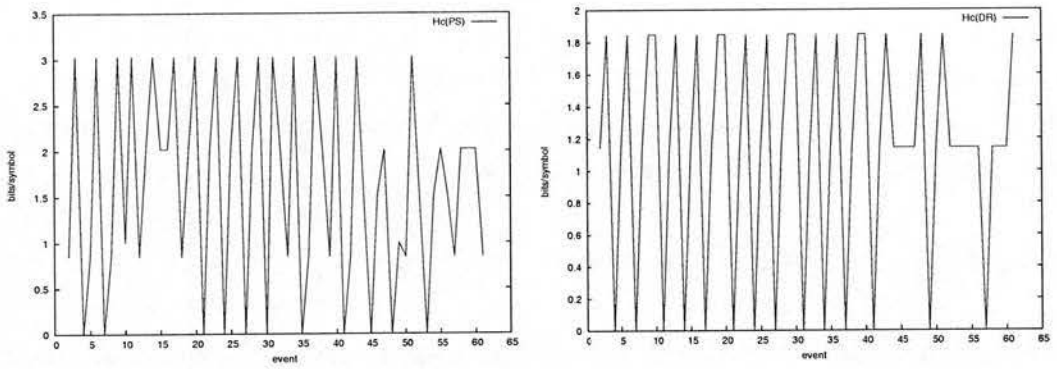
where H_{c_i} is the the entropy associated with the feature sequence (context) $c_i = (w_i, ..., w_{i-n-1})$, where n is the order of the model.

We conjecture that the negative values of the function $S(i)$, i.e. when entropy drops significantly, are predictors of the locations where segment boundaries are likely to be perceived. It follows that the strength (confidence) of the boundary predictions is in principle associated with the magnitude of $|S(i)|$.

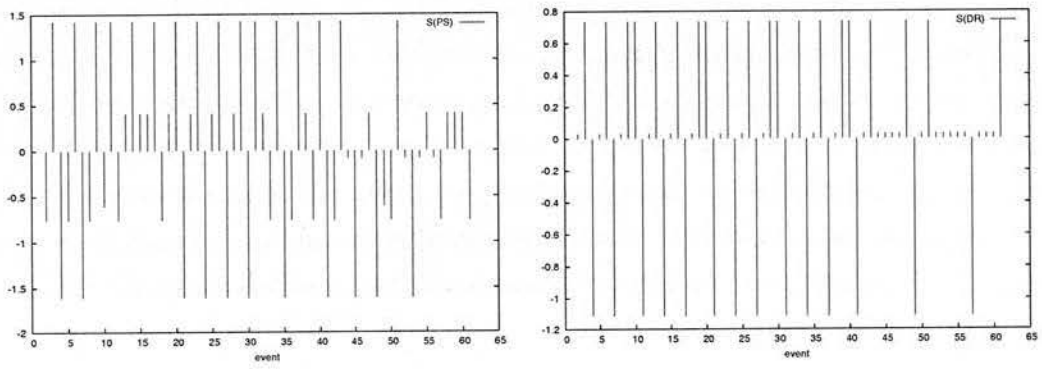
A final example for the theme of Mozart's symphony in Gm is depicted in Figure 6.4, showing the entropy profiles $H_c(PS)$ and $H_c(DR)$, and the corresponding boundary prediction profiles for $S(DR)$ and $S(DR)$. The score of this melody is also shown, with an indication of the locations of the boundary predictions. Note that for this example only the stronger predictions for each feature were selected.

6.4 Summary

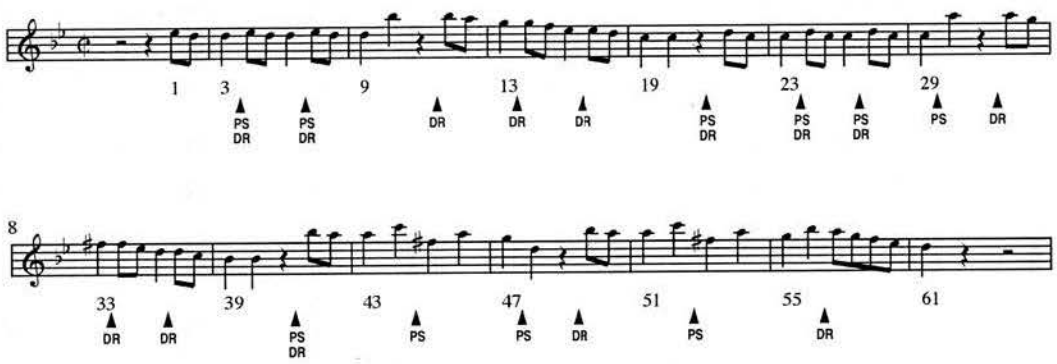
In this chapter we described in detail a melodic segmentation model, based on a probabilistic learning paradigm. It was demonstrated how this model can be trained with sequences of pitch-based and time-based melodic features (extracted from a given melody in MIDI format), in analogy with the familiarisation of listeners with a new melody. We described how estimated sequence probabilities are then used to generate entropy profiles for each one of the melodic features considered. Finally we hypothesised that segment boundaries perceived by listeners are likely to occur in the vicinity of prominent changes in these entropy profiles. In the next chapter we report some experimental results, obtained using this segmentation model to make boundary predictions for all the melodies used in the listening study (presented in the previous chapter).



(a)



(b)



(c)

Figure 6.4: Example of boundary prediction for the theme of Mozart's Symphony in Gm **a)** Entropy profiles $H_c(PS)$ and $H_c(DR)$ **b)** Boundary prediction profiles $S(PS)$ and $S(DR)$ **c)** Score with indication of boundary predictions

Chapter 7

Experimental Results

In this chapter we present the experimental results obtained with the probabilistic segmentation model described in the previous chapter. The boundary predictions generated by the model are evaluated by comparing them with the listening study presented in Chapter 5.

7.1 Methodology

To evaluate the predictions of the segmentation model we compare them with the listeners' segmentation boundaries. For this comparison to take place we first need to establish three important criteria: first, the criterion that identifies the relevant segmentation boundaries in the probability density profiles obtained from the listening experiments; second, the criterion that identifies points of outstanding entropy variation, in the entropy profiles generated by the model, for the features considered; third, the criterion that establishes if there is a correspondence between the model predictions and the listeners' segmentation data.

Selection of listener boundaries Listener boundaries (LB) are defined to occur on every local maximum of the probability density function with a peak value greater than $\frac{1}{3}$ of the maximum value.

This cut-off level was set upon analysis of the probability density graphs for all the melodies. It turns out that the majority of LB peaks that lie below $\frac{1}{3}$ of the maximum, are very single-sided in terms of subject representation, meaning that they often correspond only to the segmentation intentions of either musician or non-musician subjects. So the aim here was to select the most prominent boundaries for each melody, but also to have

balanced listener representation in the boundaries selected. It must be noted that there is not an exact correspondence between a LB peak value and the number of subjects that contributed to it. This is due to the fact that not all subjects indicated the same number of boundaries and also because boundaries that contribute to a probability density maximum are not strictly simultaneous. With our cut-off level set at $\frac{1}{3}$, no selected LB has less than 10 contributing listener votes.

From this point onwards we may refer to a LB by the index of the event that follows it, so that a LB peak in the probability density function specifies the cut-off point between segments. We consider that there is an LB before an event e_i, if a LB peak occurs between the onset times of events e_{i-1} and e_i, so event e_i is the start of a new segment. This follows a strictly temporal interpretation of the LB probability density function in comparison with the source melody events, and no assumptions are deliberately made at this stage, on if the listeners anticipated or delayed their responses.

Selection of model predictions The selection of the model's boundary predictions is based on the statistical properties of the entropy profiles H_c of each melodic feature considered. Considering the distribution of entropy around its mean value, the *stdev* can be used as a threshold to select only part of that distribution, corresponding to the greater deviations from the mean.

We have considered two threshold levels at $1\times$ and $2\times$ the $stdev(H_c)$. So for the more selective threshold, a prominent boundary prediction is defined to exist at an event e_i if $S(i) < -2 \times stdev(H_c)$, where $S(i)$ is the entropy-based feature salience function defined in Chapter 6. In the presentation of results that follows, we have focused mainly on the more selective of the two thresholds, but whenever relevant, we address weaker boundary predictions that are selected only by the lower threshold ($< 1 \times stdev(H_c)$).

We acknowledge the subjectivity of these evaluation parameters, but they are the starting point for an evaluation of our model, and they are aimed to provide a first cut on the boundary predictions. Later in this chapter we provide a comparison of the prediction results for the two considered threshold levels, to assess the impact of the adopted selection criteria and the robustness and accuracy of the model's predictions. In the next chapter we revisit this issue and discuss other possible ways of selecting boundary predictions.

As with the listener boundaries, from this point onwards we may refer to a segment boundary prediction by the index of the event that follows it. So for example, if we would refer to a boundary at event 10, it means that e_{10} establishes the start of a new

segment. We may also refer to boundaries related to a particular melodic feature, by the index of the corresponding feature vector. For example boundary $PS(10)$ would refer to a boundary prediction based on Pitch Step occuring at event e_{10}.

Boundary matching A boundary prediction at an event e is considered to be correct if e occurs within a temporal window of $\pm\tau$ around an existing LB peak, where τ is the maximum time-response of the listeners, estimated in Chapter 5. Whenever more than one event occurs whithin this time window, each one of these events is a candidate for a correct boundary prediction. In the cases where the forward time window does not overlap any event (e.g. if the LB peak occurs during a rest) then the event immediately following is also considered. We allow this matching tolerance mainly due to the fact that, as mentioned previously, there is some variance in the timing of the listeners' responses. But for this reason, the tolerance window is based on a parameter that is obtained from the listening data.

For a quantitative evaluation of the predictive performance of the segmentation model, we need to consider not only the boundaries that are correctly identified by the model, but also the listeners' boundaries that are omitted by the model, or the model predictions that have no correspondence with the listeners' melodic segments. The two measures, *Precision* and *Recall* (Van Rijsbergen, 1979), already discussed in Chapter 4, are suited for this evaluation because they focus on the positive matches, a desirable characteristic when the negative examples outnumber the positive examples. In our case, true positives (TP) are boundaries that are in both the reference (the LBs) and the predictions, false positives (FP) are boundaries that are in the predictions but not the reference, and the false negatives (FN) are boundaries in the reference not in the predictions.

Precision (P): gives the ratio of segment boundaries indicated by listeners who figure in the model's predictions. This measure reflects the boundaries that the model is not able to predict.

$$P = \frac{TP}{TP + FP} = \frac{\textit{no. boundaries correctly predicted}}{\textit{total no. LBs}} \tag{7.1}$$

Recall (R): gives the ratio of the model's predictions that appear in the listeners segmentations. This measure reflects the occurrence of incorrect boundary predictions. Incorrect or excessive boundary predictions are those that have no correspondence with any of the L-boundaries. Thus

the total number of boundaries predicted, referred in equation 7.2, includes all predictions both correct and incorrect.

$$R = \frac{TP}{TP + FN} = \frac{\textit{no. boundaries correctly predicted}}{\textit{total no. boundaries predicted}} \tag{7.2}$$

7.2 Boundary Prediction Results

We now present the results of the model's predictions for each one of the melodies used in the listening study. For each melody we present all boundary predictions in the form of a table. Each boundary prediction is specified by the number of the event that follows the boundary, the indication of bar number (for ease of reference on the score) and an indication if the boundary has a matching LB in the listening graphs (indicated with a 'y' or 'n').

In the column *Observations* we characterise some of the predictions, with an indication of the characteristics of the underlying melodic passage. We distinguish two main indications: A 'proximity' indication highlights predictions in locations that can be related to the Gestalt-based principle of proximity. This includes occurrences such as rests, events with long duration or large pitch intervals. In the particular case of Syrinx these indications will also refer to breath marks, which from the performance point of view, correspond to silences between events. A 'pattern' indication identifies boundary predictions in locations that can be associated with the reoccurrence of pitch interval or duration patterns.

For the benefit of clarity and conciseness in the presentation of these results, data charts not explicitly referred to in this section can be found in Appendix D.

7.2.1 Syrinx

In Table 7.2 we list the selected LBs for Syrinx, and in Table 7.1 we show the predictions of the model.

From Table 7.1 we obtain the following counts:

- No. of correctly predicted boundaries: 7 (PS: 4, DR:4, PS+DR:1)
- No. of incorrectly predicted boundaries: 7 (PS: 5, DR: 2)
- No. of boundaries not predicted: 5/12 (at events 42, 50, 80, 112, 300)

PS				DR			
Event	*Bar*	*LB*	*Observations*	*Event*	*Bar*	*LB*	*Observations*
26	4	n		14	2	y	pattern, precedes b.m.
29	4	n	breath mark (b.m.)	223	24	y	proximity
34	5	n		251	24	y	pattern
39	5	n		253	24	y	proximity
52	9	y	proximity, b.m.	263	27	n	follows b.m.
138-9	17	y	proximity, b.m.	268	28	n	
152	19	y	proximity				
248	24	y-	proximity				
304-5	32	n	b.m.				

Table 7.1: Characterisation of boundary predictions for Syrinx

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
13360	100	14,15
24560	33	41,42
27780	54	47,48
31910	45	49,50
36200	79	52,53
46050	68	79,80
56270	86	111,112
67590	43	138,139
86410	50	224,225
93370	72	251,252
94600	45	252,253
115670	46	299,300

Table 7.2: Listeners' boundaries selected for Syrinx. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

and the corresponding measures of Precision and Recall:

$$P = \frac{7}{12} = 0.58 \quad , \quad R = \frac{7}{7+7} = 0.50$$

There is an equal contribution of features PS and DR to the correct predictions of the model. Upon analysis, it was found that these predictions are mostly associated with discontinuities of the melodic surface, either due to longer notes or silences (e.g. DR(15,251,253)), the latter mostly due to breathing points, and large registral changes (e.g. PS(52,138,152)).

Boundary predictions PS(26,29,39,304), and DR(263), although all labelled as incorrect according to our evaluation criteria, do match weaker (but not selected) LBs, at these event locations.

We now turn our attention to some of Syrinx's LBs not predicted by the model. It is possible that some of these may be displaced due to either anticipated or delayed reaction of listeners. This seems to be the case with LB at event 50 which is likely to be an anticipated response to the long notes, just preceding the repetition of the opening motif, correctly predicted by PS(52).

Other LBs not accounted for, like those at events 42 and 80, do however match weaker predictions in the S(DR) graph, but are not selected with the higher threshold.

Because there is a significant number of boundary predictions so close to the selection threshold line, particularly visible in the S(DR) chart (Figure D.1), if we adopt the lower selection threshold ($1 \times stdev(H)$) we add also a considerable number of incorrect predictions. A total of 22 additional boundaries would be selected, all from DR. These include 6 correct predictions of which only 2 are new (DR(42) and DR(78)) with the remaining 4 overlapping previous predictions. There are 16 additional incorrect predictions, 12 of which correspond to the locations of the three-note repeating pattern of the opening motif (DR(2,5,12,16,19,26...)). If we consider these weaker predictions, and recalculate *Precision* and *Recall*, we obtain:

$$P = \frac{7+2}{12} = 0.75 \text{ (previously 0.58)} \quad , \quad R = \frac{7+2}{12+22} = 0.25 \text{ (previously 0.50)}$$

Thus the inclusion of weaker predictions although increasing *Precision* penalises *Recall* due to the additional incorrect predictions added.

Performance issues Because the Syrinx MIDI file used in these experiments corresponds to an expressive performance, some dynamic and sometimes subtle changes in

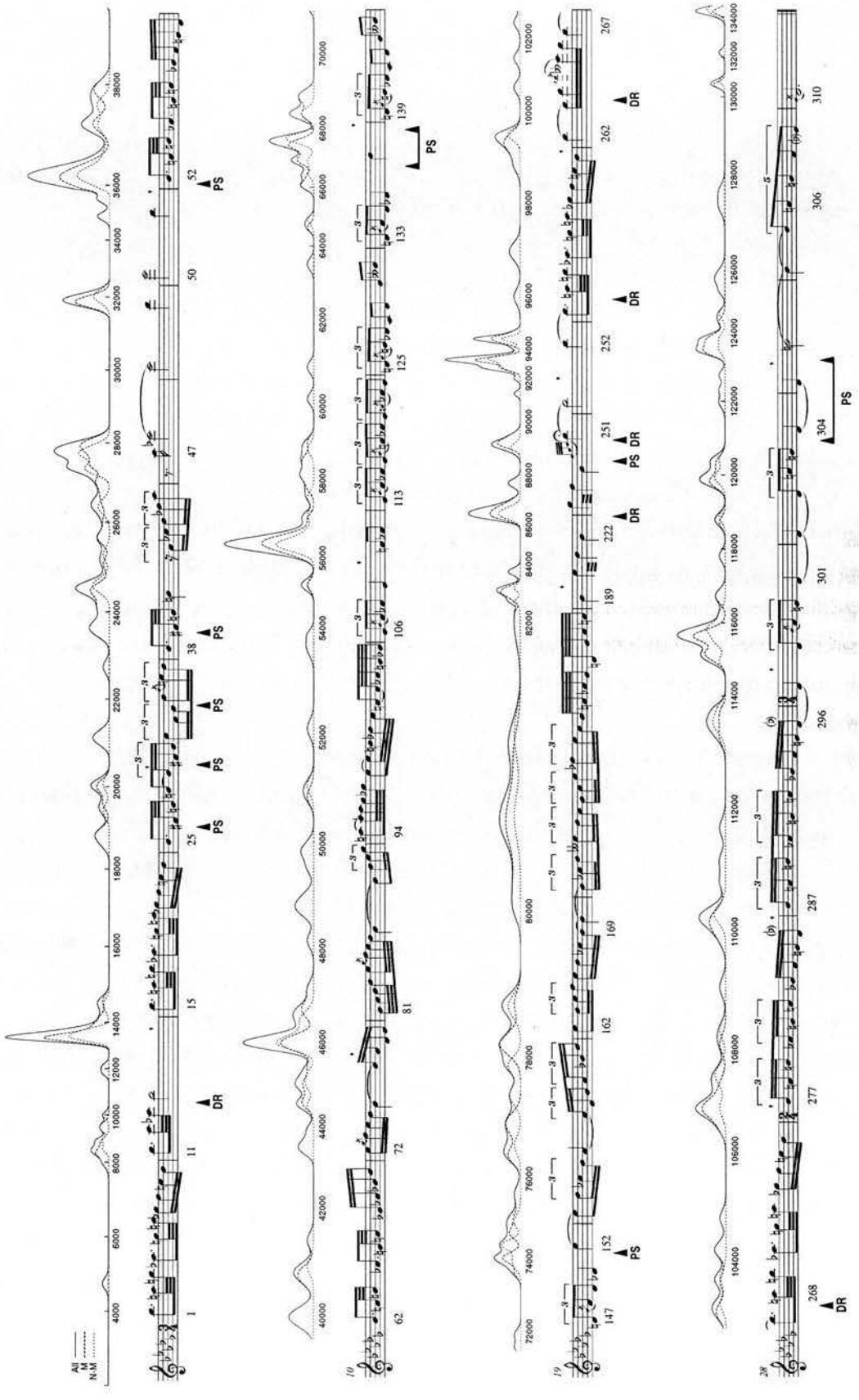


Figure 7.1: Comparison between model boundary predictions and listeners' boundaries for Syrinx

Bar no.	DR sequence
1	(-3, 0, 3, -3, 0, 1, 0, 0, 0)
3	(-3, 0, 3, -3, 0, 2, 0, 0, 0)
8	(-3, 0, 3, -3, 0, 1, 0, 0, 0)
9	(-3, 0, 3, -3, 0, 2, 0, 0, 0)

Table 7.3: Duration ratio representation for four occurrences of Syrinx' opening motif, showing slight differences due to variable event durations

the melodic features may have influenced listeners as they executed the segmentations. For example, in the vicinity of LB 112 there seems to be a strong sense of continuation, across the whole of bar 14 and then over to bar 15. It seems that the occurrence of a breath point may have influenced the listeners' boundary indication, though this was not enough to produce a significant change in the entropy profile. A similar context involves LB 300, which follows a breath point before event 297.

Other effects resulting from Syrinx's expressive melodic data include melodic patterns that were expected to be similar, but turned out to be distinct. This is true only for the DR feature where some sequences of events, although notated with identical duration on the score have in fact distinct real-time duration. In Figure 7.3 four occurrences of the same melodic sequence are shown to have slight differences when represented by duration ratios.

Breath points increase the distance between events by introducing silences. Since breath marks tend to be placed (and performed) at the start of phrases, the association between event onset distance and phrase boundaries is strengthened. There is a significant overlap between breath marks and LBs, which can be observed in Figure 7.1 and as expected, some predictions also indicate some of these locations.

All performed ornaments are represented in the midi file as individual events, as shown in the example of Figure 7.2). Some ornaments may affect the statistics of the melodic features typically increasing the counts of small pitch and time-based intervals. In order to find out if this influenced the results we edited the original Syrinx MIDI file, removing all ornaments. Then we used the edited file to train the model and generated new boundary predictions. We observed that overall, results remained unchanged with the exception of boundary PS(248), which no longer appears in the predictions since e_{248} is itself a grace-note. Also boundary PS(138-139) has gained definition and appears now only as PS(139). The two trills in bars 23 and 24 are responsible for the majority of ornamentation events, introducing sequences of alternating ascending/descending

major and minor seconds and rhythmically, sequences of identical durations.

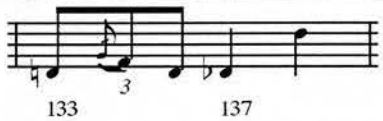
Example 1: Grace-note in Bar 16	Example 2: Trill in Bar 23
id, pitch, onset, duration 133, 62, 63912 508 **134, 66, 64420 156** 135, 65, 64576 528 136, 62, 65104 392 137, 61, 65496 1564 138, 73, 67060 1142	id, pitch, onset, duration 189, 70, 82269, 579 **190, 75, 82848, 70** **191, 73, 82918, 71** **192, 75, 82989, 71** **193, 73, 83060, 70** ... **220, 75, 84948, 71** **221, 73, 85019, 69** 222, 70, 85088, 1190
 133 137	 189 222

Figure 7.2: Examples of event sequences representing ornaments from the Syrinx MIDI file

Although these results show that the boundary predictions for Syrinx are somehow 'immune' to the presence of ornaments, it is likely that other highly ornamented pieces might be more affected by the additional number of events. Partly due to its length, Syrinx can accommodate the extra events without altering the salience of other relevant features.

7.2.2 Sonata K284

Boundary predictions for melody K284 are shown in Table 7.4. It can be seen that all except two PS-based predictions have corresponding LBs. Predictions for this melody seem to be predominantly associated with the occurrence of either large pitch intervals or temporal distance between events. In the latter cases, distances between some events have been accentuated by the presence of rests, which occur at the end of some melodic phrases. There are also three LBs that are correctly indicated by overlapping PS and DR-based predictions.

The two incorrect predictions (PS(37) and PS(87)) result from two occurrences of the same melodic passage in measures 7 and 15, but have little or no expression in the listener's data.

The only LB not predicted by the model occurs at event 51. Here, the absence of either pitch or duration discontinuities suggests that maybe only a combination of features

PS				DR			
Event	*Bar*	*LB*	*Observations*	*Event*	*Bar*	*LB*	*Observations*
26	5	y	prox./patt.	101	17	y	pattern
37	7	n		127	22	y	prox./patt., rest
76	13	y	same as 26	152	10	y	pattern
87	15	n		178	31	y	prox./patt., rest
127	22	y	prox./patt.				
152	26	y	prox./patt.				
178	31	y	same as 127				

Table 7.4: Characterisation of boundary predictions for melody K284

could predict the boundary at this point, by capturing the reoccurrence of the melody's opening 3-event pattern.

From Table 7.4 we obtain:

- No. of correctly predicted boundaries: 6 (PS: 5, DR: 4, PS+DR: 3)
- No. of incorrectly predicted boundaries: 2 (PS: 2, DR: 0)
- No. of boundaries not predicted: 1/7 (at event 51)

and the corresponding measures of Precision and Recall:

$$P = \frac{6}{6+1} = 0.86 \quad , \quad R = \frac{6}{6+2} = 0.75$$

7.2.3 Sonata K333

All prominent LBs for melody K333 are predicted by the model and correspond essentially to DR-based predictions, as shown in Table 7.5. It can be observed that these time-based predictions are associated either with rests or prolonged notes. The boundary prediction at event 128, the only in the DR set not identified by listeners, coincides with an alteration in a rhythmic pattern that reccur at the start of measures 2 and 12.

At the pitch level, many of the predictions are associated with large intervals that contrast with the scale-step sequences that are frequent throughout the melody. An example of this includes the three erroneous boundary predictions at events 184, 188 and 192 (measure 20), which correspond in fact to a sequence of notes distributed between different voices, with alternating melodic and harmonic function.

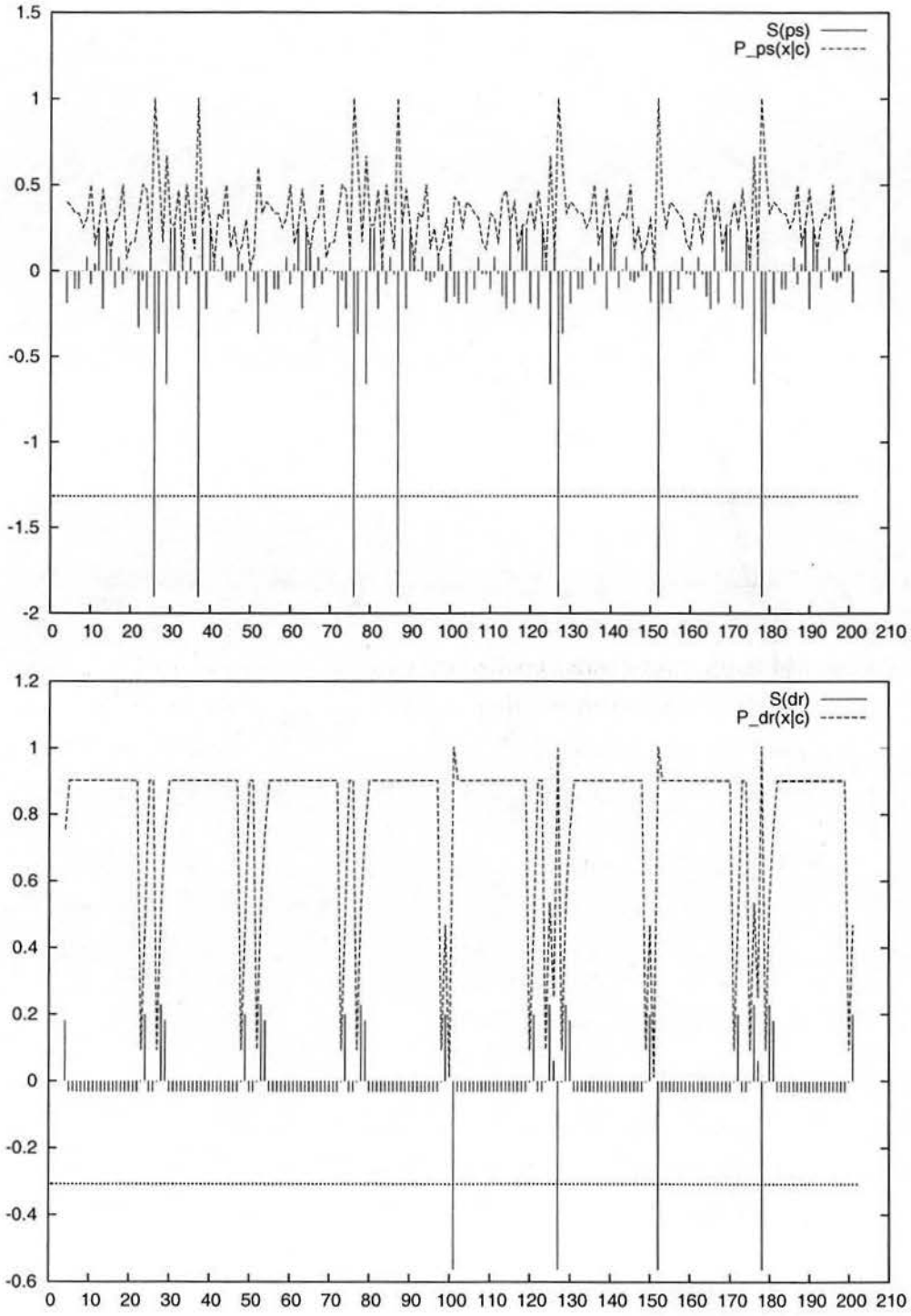


Figure 7.3: Sonata K284: boundary predictions $S(c_i)$ and successor probability $P(x|c_i)$ for features PS and DC. Boundary threshold indicated by dotted line at bottom of the graph

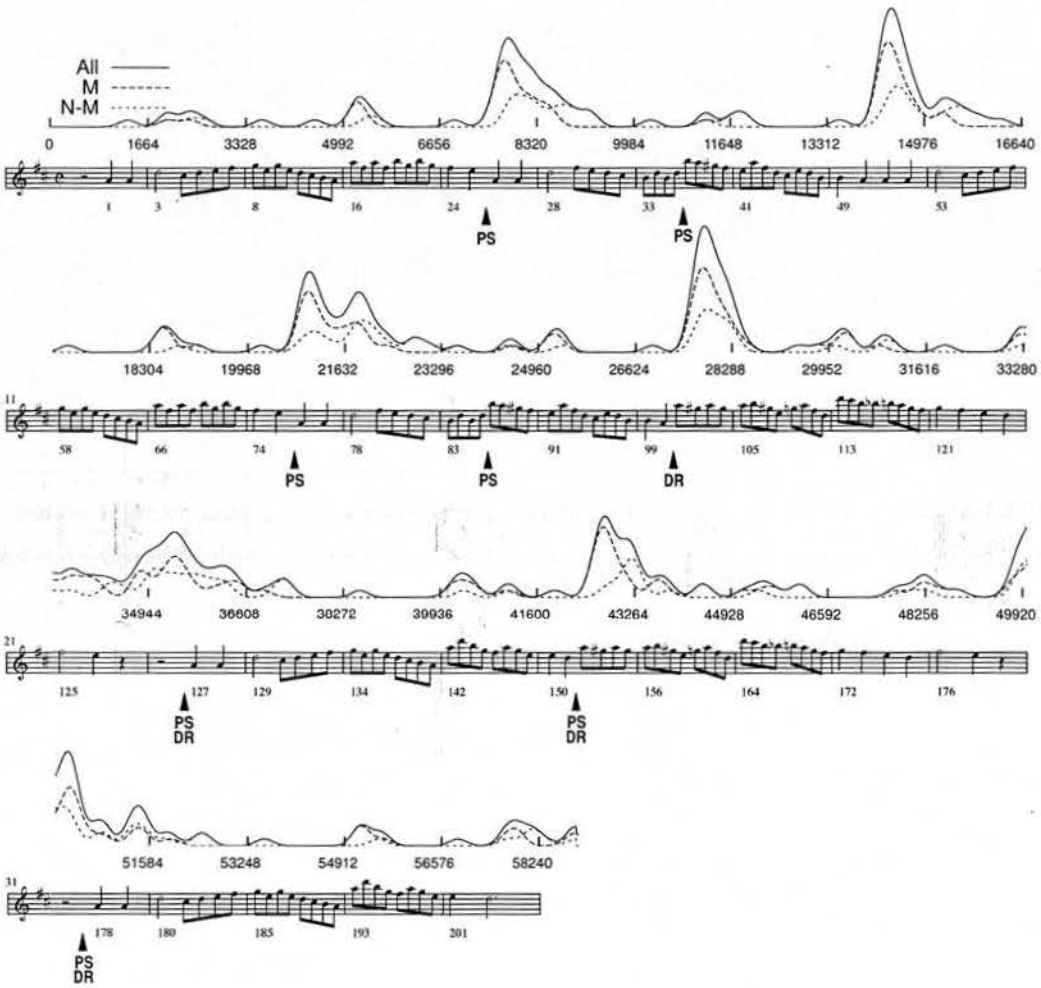


Figure 7.4: Comparison between model's boundary predictions and listeners' boundaries for melody k284

PS				DR			
Event	*Bar*	*LB*	*Observations*	*Event*	*Bar*	*LB*	*Observations*
83	10	n		13	2	y	proximity
91	11	n		26	4	y	proximity, rest
109	13	n		84	10	y	prox./patt, rest
171	19	n		124	14	y	prox./patt, rest
180	20	y	proximity	128	15	n	
184	20	n					
188	20	n					
192	20	n					
202	22	n					

Table 7.5: Characterisation of boundary predictions for melody K333

We have found that boundary predictions PS(91) and PS(171), although marked as incorrect, have a correspondence with two weaker LBs. Prediction PS(83) coincides with a descending octave which marks the end of a melodic phrase (see Figure D.3). From a perceptual point of view, this is arguably a significant prediction but it is temporally too distant to the neighbouring LB (occuring at event 86) to be matched as correct, according to our established evaluation criterion.

So from Table 7.5 we obtain:

- No. of correctly predicted boundaries: 5 (PS: 1, DR:4)
- No. of incorrectly predicted boundaries: 8 (PS: 8, DR: 1)
- No. of LBs not predicted: 2/7 (at events 76, 97)

and the corresponding measures of Precision and Recall:

$$P = \frac{5}{7} = 0.71 \quad , \quad R = \frac{5}{5+9} = 0.36$$

Although more than half the LBs are predicted by the model the proportionally large number of incorrect predictions is responsible for the low *Recall*. If we consider the less selective threshold, all LBs are accounted for but additional incorrect predictions lower the *Recall* to 0.28.

It is worth noting that the two LBs not predicted by the model while using the higher threshold are in fact the weaker of the set. It is also interesting to observe that one of these (76) is indicated exclusively by musician subjects, so arguably it does not represent the intentions of all of the listeners.

DR			
Event	*Bar*	*LB*	*Observations*
5	2	n	
11	4	y	proximity
16	6	n	
22	8	y	proximity
27	10	n	
33	12	n	

Table 7.6: Characterisation of boundary predictions for melody E0547

7.2.4 Folk-song E0547

From Table 7.6 we observe that only the DR feature provides predictions for melody e0547. Only two of the four LBs considered are correctly predicted by the model. The majority of boundaries predicted from DR have little correspondence with the listening data. It appears that the small size of the training data (only 39 events) had an influence on the relative probability of some event durations. This is particularly visible in the graph of Figure 7.5 where $H(DR)$ is seen to oscillate continuously between only three different entropy levels.

From Table 7.6 we obtain:

- No. of correctly predicted boundaries: 2 (from DR)
- No. of incorrectly predicted boundaries: 4
- No. of boundaries not predicted: 2/4 (at events 30 and 35)

and the corresponding measures of Precision and Recall:

$$P = \frac{2}{4} = 0.50 \quad , \quad R = \frac{2}{2+4} = 0.33$$

It is worth mentioning that although boundary DR(22) is identical to boundary DR(11), the former has a weaker correspondence in the listeners' graph. This apparent inconsistency, already discussed in Chapter 5, highlights the difficulties of comparing model predictions and real listening data. This issue will be revisited in the next chapter.

From Figure 7.6 it can be seen that some PS-based predictions lie just above the selection line. If we adopt the lower selection threshold, we add six additional incorrect predictions, and as a result we lower the value of *Recall*. These boundaries have been

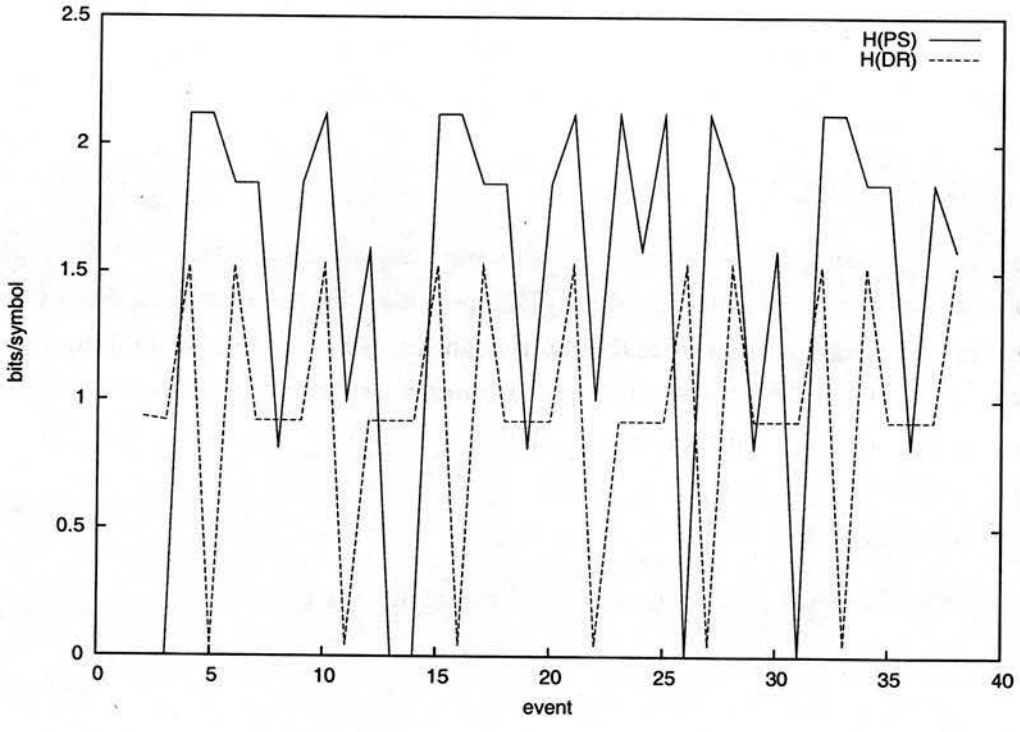


Figure 7.5: Entropy profiles $H_c(PS)$ and $H_c(DR)$ for folk-song E0547

DR			
Event	*Bar*	*LB*	*Observations*
2	1	n	
14	3	y	proximity

Table 7.7: Characterisation of boundary predictions for melody f0927

indicated in parenthesis in Figure D.4 for visual reference. It seems that the small number of events of this melody and the predominance of intervals of a major 2nd makes all other intervals statistically very prominent. This is the case of intervals of a minor 2nd which are mostly responsible for four of these PS-based boundaries.

The two LBs not predicted by the model are the weaker of the four.

7.2.5 Folk-song F0927

Only the DR feature generates predictions, using the most selective threshold. The graph for the DR predictions, depicted in Figure D.5, provides a good illustration of a case where average entropy is low due to limited rhythmic diversity. In this cases, boundary salience, as we have defined it, will be very sensitive to any slight changes in the duration ratios. Nevertheless some structure is still visible in these graphs, and one correct boundary prediction was obtained.

From Table 7.7 we obtain:

- No. of correctly predicted boundaries: 1 (from DR)
- No. of incorrectly predicted boundaries: 1
- No. of boundaries not predicted: 2/3 (at events 26 and 38)

and the corresponding measures of Precision and Recall:

$$P = \frac{1}{3} = 0.33 \quad , \quad R = \frac{1}{1+1} = 0.50$$

By adopting the lower selection threshold we observe (see Figure D.5) that 10 PS-based additional predictions are selected. These predictions (indicated in parenthesis in Figure D.6) locate the two previously missing LBs, at events 26 and 38, and two others (PS(7) and PS(31)) also match weaker LBs. Halving the selection threshold also adds 4 incorrect predictions to the results, so the new values of *Precision* and *Recall* would

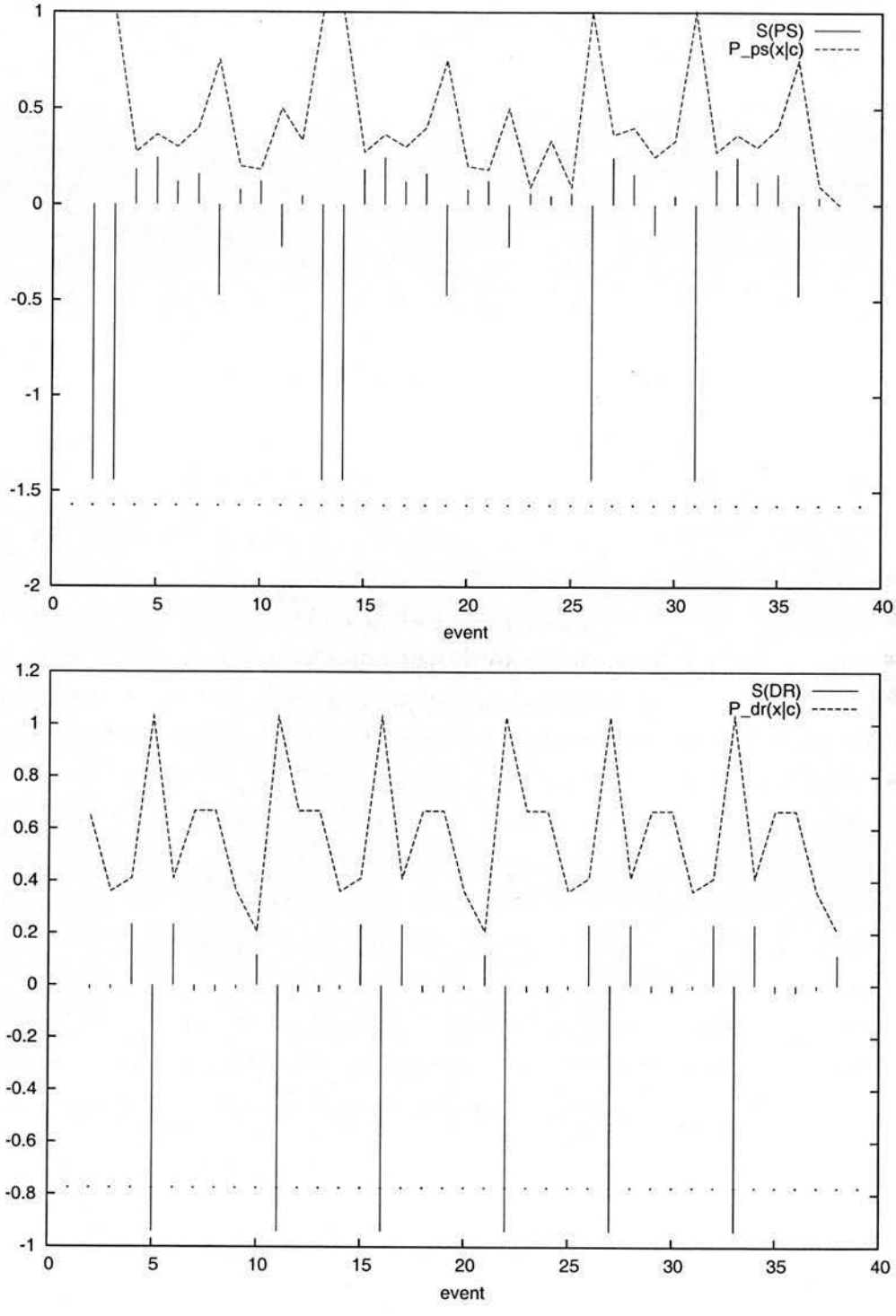


Figure 7.6: Folksong E0547 boundary predictions, showing boundary salience $S(c_i)$ and successor probability $P(x|c_i)$ for features PS and DC. Boundary selection threshold indicated by the dotted line at bottom of the graph

PS				DR			
Event	*Bar*	*LB*	*Observations*	*Event*	*Bar*	*LB*	*Observations*
3	2	n		9	3	n	
16	5	n		12	4	n	
29	8	n		14	4	y	prox./patt.; rest
34	8	n		22	6	n	
53	12	y	prox./patt.	25	7	n	
				27	7	y	proximity; rest
				37	9	y	proximity
				46	11	y	proximity
				53	12	y	prox./patt.
				56	13	n	

Table 7.8: Characterisation of boundary predictions for melody Q0034

then be: :

$$P = \frac{5}{5} = 1.00 \quad , \quad R = \frac{5}{5+5} = 0.50$$

7.2.6 Folk-song Q0034

In Table 7.8 it can be seen that all listeners boundaries are predicted by the model, but at the cost of a proportionally high number of incorrect predictions. Mostly DR contributes with accurate boundary predictions, and only one of the PS boundaries has correspondence in the listeners' segmentations.

Boundary predictions PS(3,16,53) are associated with an identical pitch sequence that starts with an ascending fifth, although only PS(53) is supported by the listeners data. Note also that boundary PS(29), marked as incorrect, is not only melodically plausible (e.g. events 27 and 28 perceived as an *anacrusis*), but it finds some support in the listeners' segmentations (see Figure D.7), where a prolongation of the LB peak is visible, past the end of bar 7, suggesting that some listeners might have indicated a boundary closer to the start of bar 8.

From Table 7.8 we obtain:

- No. of correctly predicted boundaries: 5 (PS: 1, DR:5, PS+DR: 1)
- No. of incorrectly predicted boundaries: 8 (PS: 4, DR: 4)
- No. of boundaries not predicted: 0/5

and the corresponding measures of Precision and Recall:

$$P = \frac{5}{5+0} = 1.00 \quad , \quad R = \frac{5}{5+8} = 0.38$$

In Figure D.9 we observe that all selected boundary predictions have the same strength, both for PS and DR. This means that no discrimination is possible by adjusting the selection threshold alone. The values of *Precision* and *Recall* are therefore not affected if the lower selection threshold is used.

7.3 Discussion of results

Table 7.9 summarises the predicted results for all six melodies, showing the values of *Precision* and *Recall* for the two selection threshold levels considered. For comparison, this table also indicates the length of each piece.

With the more selective threshold, the model has an average *Precision* of 0.66, meaning that overall more than half the listeners' boundaries were correctly predicted. The corresponding average *Recall* of 0.47, indicates that the model generated approximately one incorrect prediction for each correct one. As we might expect a trade-off between higher *Precision* and lower *Recall* is obtained by relaxing the selection threshold. This trade-off is more apparent for Syrinx and the two Mozart sonata excerpts, which are also the three larger member of the melody set. For the remaining three smaller pieces, the less selective threshold practically did not alter the prediction rates. On one hand, this shows that the selection process is robust to changes in the selection threshold, but it also reflects a lack of diversity in terms of boundary salience, which is more striking in the smaller pieces. In some cases all salient boundary predictions have identical values, so they are either all selected or all rejected. This is particularly visible in the $S(DR)$ prediction graphs, and may result from the fact that DR is a more abstract melodic representation than PS. Because DR has a considerably smaller alphabet of possible symbols, the range of values of the entropy function is more limited. Note that for most melodies, the range of values contained in the corresponding features is usually a sub-set of the full range defined for DR and PS.

In Table 7.10 we provide a characterisation of the model's correct boundary predictions by showing the different contribution of the two melodic features PS and DR. The table shows that DR is the predominant melodic feature amongst the correct predictions with 20 boundaries, against only 11 from PS.

Melody	*No. events*	*Duration (sec.)*	$2 \times stdev(H)$		$1 \times stdev(H)$	
			Precision	*Recall*	*Precision*	*Recall*
Syrinx	311	133.9	0.58	0.50	0.75	0.25
K284	202	57.5	0.86	0.75	0.86	0.50
K333	209	45.5	0.71	0.36	1.00	0.28
E0547	39	16.8	0.50	0.33	0.50	0.33
F0927	45	15.6	0.33	0.50	1.00	0.50
Q0034	57	19.8	1.00	0.38	1.00	0.38
		Averages	0.66	0.47	0.85	0.37

Table 7.9: Summary of boundary prediction results, for the two threshold levels considered.

This predominance of DR-based predictions may be the result of several factors. If, as we noted in Chapter 5, there is a preponderance of LBs coinciding with rests or long notes, then it would be expected that a time-based feature is more able to acquire a larger portion of these occurrences. Also since DR is a more abstract representation than PS, the latter is more likely to suffer from data sparseness, and therefore more likely to generate predictions which are too specific or too localised. Duration ratios benefit from higher redundancy for the same amount of training data, so in principle, predictions are based on better probability estimations.

Table 7.10 also shows that the great majority of correct boundary predictions occur in locations that can be associated with the Gestalt-based principle of proximity. Note that these numbers include both pitch and time based occurrences. Fewer predictions appear in locations associated only with pattern recurrence.

On one hand, this predominance of boundary predictions based on surface discontinuities may be related to the fact that listeners are significantly influenced by surface discontinuities, much in the same way that DR-based predictions were dominant amongst LBs. This hypothesis finds support in previous empirical studies (Deliège, 1987; Deliège and Melén, 1997; Deliège, 1998) where it was reported that listeners' decisions were dominated by Gestalt principles of proximity.

Another important factor relates to the fact that to express surface discontinuities we require less stored context than we need to express most parallel melodic passages. A large pitch interval or a large duration ratio can be obtained with only two events, whereas in most cases patterns obtained from fewer than three events may be considered too general. The resulting low order of the models trained with our melodic features, which lie somewhere between a bigram and trigram models, may be limiting the ability

Melody	*No. correct predictions*	Source		Type	
		PS	*DR*	*Prox.*	*Patt.*
Syrinx	7 (11)	4 (4)	4 (8)	6 (12)	2 (4)
K284	6 (6)	5 (5)	4 (4)	5 (5)	6 (6)
K333	5 (7)	1 (3)	4 (6)	5 (5)	2 (4)
E0547	2 (2)	0 (0)	2 (2)	2 (2)	0 (0)
F0927	1 (5)	0 (5)	1 (1)	1 (5)	0 (4)
Q0034	5 (5)	1 (1)	5 (5)	5 (5)	3 (3)
Totals	26 (36)	11 (18)	20 (26)	24 (34)	13 (21)

Table 7.10: Characterisation of correct boundary predictions according to source melodic feature and type of underlying melodic principle. Boundary counts in parentesis correspond to the lower selection threshold. Numbers include overlapping predictions.

to capture parallel patterns.

It must be made clear that the 'proximity' and 'pattern' indications result from an external (and subjective) analysis of the locations in the melody, where the model predictions occur. From a probabilistic perspective the distinction between the two is not so sharp. For example, a large pitch interval between two events will be statistically salient if most intervals in the melody follow a stepwise motion. In that context, one can say that the interval is salient because a small interval followed by a large one is an improbable pattern. But statistically it is also possible to find an improbable sequence of events that has no significant surface discontinuities. The purpose of the analysis in Table 7.10 was to show how these two phenomena are represented in the prediction results of our probabilistic model.

7.4 Summary

A quantitative evaluation of the model's boundary predictions has been presented. This required the definition of subjective selection and matching criteria, which were set to measure the predictive power of the model when comparing the more salient boundary predictions with the more prominent listeners' boundaries.

The results so far, seem to support our main hypothesis, showing that prediction entropy changes, associated with certain melodic features, have a significant predictive power concerning the locations of segmentation boundaries in melodies, as perceived by listeners. Most correct predictions were found to correspond to Gestalt principles, particularly associated with time-based event information. More interestingly some non-

Gestalt related boundaries predicted by the model were also found to have a correspondence in the listeners' segmentation data.

It was shown that amongst the predictions labelled as incorrect, some actually had some correspondence (although weaker) with the listeners' data. Other predictions presented some kind of structural relevance, e.g. coinciding with melodic sub-phrasing, and arguably, could be seen as perceptually pertinent, although some of these melodic sub-divisions were not clearly expressed by a majority of listeners.

Expectation is an important, although not the sole factor in the perception of melodic boundaries. As such, we anticipated that our probabilistic model could not account for all the listeners' boundaries. Under this premise, it is remarkable that a fairly simple probabilistic learning paradigm can account for more than half the listeners' segmentation boundaries.

In the next chapter we discuss the characteristics and the limitations of this model as well as some methodological options. We also re-visit some of the proposed aims and research questions in the light of our findings and other related research work.

Chapter 8

Discussion and Related Work

In this chapter we provide a critical analysis of our approach and a general discussion of our findings. We address some of the characteristics of the probabilistic model of melodic segmentation, identifying its strengths and weaknesses and relating it with existing research work.

8.1 Methodological Considerations

One of the aims of this research is to model listening behaviour, so we have argued that segmentation results obtained with the model had to be compared with real listening data. The collection of such data is time and resource consuming and, as a result, imposes a limit on the amount of melodic material that can be tested in a limited amount of time. Questions brought to light during the analysis of the results, would have benefited from additional experiments, to test other melodic samples with particular characteristics. Some of these questions are presented here, as they are a valid part of these research results, and will inform future work in this area.

The empirical study presented in Chapter 5 showed that data obtained in real-time can be challenging to interpret and use. Care was taken to use this data with minimum interference, as often one is tempted to make assumptions about the listeners' intentions, based on our perception and analysis of the melodies.

Instead, we opted to reflect some of the characteristics of the data into our evaluation methodology. The definition of subjective matching criteria was necessary to establish a correspondence between the listening data, continuous in nature, with the model's predictions, discrete in nature.

For example, a tolerance window, based on estimates of the response times of the listeners, was used to match the boundary predictions of the model with the listener's boundaries (LBs). This tolerance was allowed both before and after the location LB probability maxima, so it is assumed that LBs may be indicated with delay or anticipation. Intuitively, it would seem that in a real-time listening experiment segmentation data would be more commonly affected by delayed responses. However, our empirical study shows that both types of responses seem to be present in the data and in some cases, coexist within the same melody. An observable (although overall not statistically significant) difference between these two types of responses was found to be related with the subjects' musical knowledge, as musically trained listeners revealed higher anticipation on their segmentation responses.

From an experimental point of view it is difficult, if not impossible, to represent the (musical) knowledge of listeners accurately, prior to a new auditory experience. The need for a high number of participants in a listening experience and the separation of these subjects in two groups with distinct musical training and experience, is a methodological necessity, and an attempt to address the musical knowledge factor. Overall, the significant correlation of the listeners' boundary profiles has allowed us to consider the sum of all boundary contributions. This correlation is in itself an interesting result, as it suggests that melodic segmentation may rely on perceptual/cognitive abilities shared by all subjects and perhaps not specifically acquired through formal musical training.

8.2 The Influence of the Representation

For most computational models of music, the choice of a representation is quite influential on the results that are produced and our models are no exception. Abdallah and Plumbley (1999) criticise the fact that often the choice of a representation is made heuristically, based on intuitions of what makes sense musically in a given context.

In our experiments, MIDI files are the source of melodic information. The use of artificial musical stimuli in a listening study (such as dead-pan MIDI files) restricts the amount of auditory information available to the listener and may exclude important aspects of a real performance. Although it can be argued that these melodic samples are not fully representative of a real musical experience, their audition does constitute a musical experience of some kind, and it is this auditory experience that is the subject of our study.

Since the input to our model is a MIDI-style event-based representation, already there is a restriction on what melodic attributes are made available. Some musical dimensions such as timbre and dynamics for example, are not present in our melodic samples, and all melodic events have quantised pitch and with the exception of Syrinx, quantised durations.

As a starting point, we have chosen two separate melodic features based on pitch intervals and event durations. There was a deliberate attempt to base this choice on available empirical research on what features have been found to have a more significant interplay in the perception/recognition of melodies (Dowling, 1994; Platinga and Trainor, 2005). For example, the adoption of relative attributes (i.e. intervals and duration ratios) rather than absolute ones (i.e. pitch and duration) is supported by empirical findings that reveal that similarity judgments have been found to be fairly independent of musical training and primarily based on 'surface' features such as articulation and contour rather than on 'deeper' features such as motivic or harmonic relationships (Lamont and Dibben, 2001). But even assuming that it is possible to identify different sensitivities to different musical attributes, one wonders if that is sufficient to allow more accurate computational models of music perception and cognition. Is it reasonable to combine them by simply adding their individual results or instead is it necessary to consider their product as a combined feature?

Some have argued that, ideally, a model of music perception should aim to derive auditory precepts from raw audio signals, using redundancy reduction techniques, allowing the model to choose its own features (Abdallah and Plumbley, 1999). This is a valid observation and could be pursued as a future development. However this does not invalidate our approach where we have looked at percepts which were hypothesised to be important and tested that hypothesis by predicting listening behaviour.

Conklin and Witten (1995) proposed a framework with combined features ('viewpoints') and evaluate them based on the average entropy of the resulting model, thus providing a way of choosing the mix of viewpoints that better describes a given corpus of music. This approach is suitable for large training sets, but in our case, because models are generated from a single melody, combining features would aggravate data sparseness.

Already our boundary prediction results reflect some differences between the two features considered, due partly to the fact that duration ratios are a more abstract melodic representation than pitch intervals. A more abstract melodic feature (i.e. a smaller alphabet of symbols) will yield higher redundancy and therefore better prob-

abilistic estimation. On the other hand very abstract features will tend to make very general predictions, particularly if the order of the model is low. This becomes apparent in the entropy profiles based on duration ratios, where we observe a reduced range of boundary salience values, compared to the pitch based profiles.

Most music comprises several dimensions combining melody, harmony and rhythm over time together with the addition of timbre and dynamics. Aiello (1994) notes that given the richness of the musical stimulus, listening implies choosing which elements to attend to. Attentional issues in music perception are a complex matter and it was not an aim of this work to either address their theoretical implications or to incorporate these in our model. Nevertheless this suggests that the choice of relevant features must be a dynamic process rather than a pre-defined weighting of melodic attributes, as it was the case in the MDSM, the first of the segmentation models presented here. We argue that perhaps more important than finding the right combination of different features, is having the ability to measure the relevance of a feature in the context of each particular melody or even over the course of a specific musical passage. The information theoretic approach used to measure feature salience in our second model, is a step in this direction.

Some important temporal music aspects of music such as tempo and musical meter have not been addressed in our models due to time restrictions. Musical meter is a part of rhythm that is almost always present in music. A considerable amount of research has been produced for studying and modelling musical meter (Longuet-Higgins and Lee, 1982; Povel and Essens, 1985; Desain and Honing, 1999) so these aspects should not be left out of our discussion. Intuitively, the perception of metrical levels in a melody, is likely to condition the perception of phrasing and thus the location of segment boundaries. Given its complexity, the inclusion of musical meter in our model would require more than just the addition of another musical feature. In fact, research suggests that listeners combine multiple melodic and temporal features to perceive musical meter (Hannon et al., 2004). But a component of meter information would be likely to enhance not just the results, but also the data acquisition process of our probabilistic model. Studies have shown that rhythmic regularity may enhance learning and the recollection of auditory patterns (Povel and Essens, 1985; Hannon and Johnson, 2005) thus improving other processes which benefit from a more efficient and structured processing of musical information, as is the case with segmentation.

8.3 A Probabilistic Memory Model

Markov models are often criticised for their underlying assumption that an event depends only on previous events. This assumption seems to be over-simplistic if we are analysing musical sequences, however it is known (Eysenck and Keane, 1995) that human memory limitations impose a limit on our ability to establish large-span temporal relations. Several researchers have supported the idea that listeners acquire musical information by focusing on localised zones of musical piece, and that these perceptions are affected by a local context (Meyer, 1956; Narmour, 1992; Bigand, 1993).

When training the models we allowed high order dependencies (up to order ten) to be stored. However the estimated weighting coefficients of the mixed-memory model, reveal that for our corpus of music, the model can effectively only represent dependencies that are equivalent to a full-memory model somewhere between a bigram and a trigram.

Despite the resulting low order of the models, a significant number of perceptually pertinent melodic boundaries were successfully predicted. The characteristic low order of the models indicates that the patterns that are acquired are fairly low level, but they may be sufficient to detect intra-opus structural elements. We anticipate that higher level inter-opus schematic similarities may be more difficult to capture with this type of model.

Conklin and Witten (1995) proposed a combination of short and long-term models, where the long-term model represents the general musical style and the short-term model the details of an individual piece. This combined framework was used to make predictions about the information content of Bach chorale melodies , according to the authors, with reasonable success. Although not much detail is provided about the actual interaction between long and short-term components underlying the prediction results, Conklin and Witten underline the need to investigate different model combination schemes.

Other researchers have reported limitations in establishing inter-opus similarity relations with low-level melodic features. "Although some evidence exists that frequencies of events may be useful as cues for similarity, the raw frequency counts of tones, intervals, and durations and their first-order transitions were not particularly effective predictors of melodic similarity" Eerola et al. (2002, p. 286).

Critics of the probabilistic learning approach may argue that it is unrealistic to propose that listeners are able to memorise and remember a large number of patterns and

corresponding probabilities of occurrence. It is known that a piece of music conveys more information than a listener can process in a single hearing. However, retention of melodic features is enhanced by repeated hearing of a piece (Deliège and Melén, 1997; Deliège, 1998), hence the need for the familiarisation auditions in our listening study.

One of the reasons for the choice of a mixed-memory model was to overcome data sparseness. Full-memory models have the advantage of explicitly storing larger contexts, but that can be a problem when the training corpus is small. Full-memory models often have to use interpolation to calculate sequence probabilities, based on a weighted combination of n-grams of lower order (see Chapter 3). Preliminary experiments carried out with n-gram smoothing showed that when these weights are inferred from the training data, higher order components turn out to have a very small contribution to the sum of probabilities, and effectively we end up with a model that does not represent the full nth order.

The memory-dependant nature of context-based predictions poses difficult questions regarding the interpretation of the generated boundary predictions. Are expectation-driven melodic boundaries perceived in retrospect? Expectations are associated with a melodic context which is formed up to a certain moment in time, but the confirmation or denial of these expectations occurs only moments later, after new stimuli follow that particular context. So, hypothetically, the perception of change, although having its starting point in the initial melodic context, only materialises over the next few events, casting doubts over to where in the melody the 'perceived' boundary should be assigned.

8.4 Modelling similarity

With a probabilistic learning model, sequences of melodic feature data are transformed into probabilistic dependencies. The model can reflect the recurrence of similar melodic segments but is unable to recover the uniqueness of different patterns which might have identical probability of occurrence. Although similarity is implied between different (and distant) melodic segments, the model is not capable of establishing relationships between these segments. To achieve this, an explicit categorisation procedure, such as the one proposed by Cambouropoulos (1998), would be necessary. A simple hierarchic classification of segment boundary predictions would be possible though, if we rated them in terms of their salience. Our comparative analysis, using two different selection thresholds, revealed that the boundary salience of many of the predicted boundaries had similar correspondence in strength with the listeners' segmentation.

As mentioned earlier, the resulting low order of the mixture-models means that they can only acquire patterns that correspond to sequences of 4 events in length, at the most. Although these lower order models cannot represent large patterns, smaller partial patterns were acquired and results showed that some of these are indexes for longer patterns matching the beginning of larger segments. This is in tune with previous findings (Deliège and Melén, 1997) that shown that patterns are often remembered by listeners as a result of the repetition of smaller cells, often their initial section.

Some form of approximate similarity is implied by our melodic representation, as it is the case of chromatic transposition or proportional rhythmic replication. The melody extract from sonata K333 provides an example of a piece where melodic motifs are repeated under more complex rhythmic transformations. This type of parallelism is difficult to capture with our representation, and this was reflected in a few incorrect boundary predictions, for this particular melody. Note that, although these similarities were not detected as such, some of these 'unusual' pitch sequences were highlighted by the model in the form of boundary predictions. One could argue that some of the sections of the K333 sonata extract are not exclusively melodic, since notes are in fact split between voices, alternating between their melodic and harmonic function. But it is equally arguable whether one can ever find a sequence of notes that can be considered or perceived as purely melodic.

Meyer argues that the implicative effect of repetition is context dependant. He writes that "if a reiterated pattern is understood to be part of an ostinato or a ground bass, we do not necessarily expect change. Similarly, repetition in a coda or of a cadential figure repeated as an echo, has quite different effect from repetition which is understood to be part of an on-going process." (Meyer, 1973, p. 51). Some of these contextual differences may be difficult to address with a probabilistic model, particularly if our aim is to process non-annotated musical data. For example, in Chapter 7, it was shown that ornaments can remain undistinguished in a sequence of melodic events.

Overall, it seems to be the case that a probabilistic model performs better for melodies with more internal consistency than for pieces which register constant internal change. In other words, for similarity to be established it is necessary that a significant amount of repetition takes place to allow these pattern cues to be statistically salient.

8.5 Feature Learning and Gestalt

From the analysis of the results in Chapter 7 it is apparent that the majority of correct boundary predictions can be associated with Gestalt-based grouping rules. A preponderance of these grouping principles relating to listeners' boundary preferences has been reported in other segmentation experiments (Deliège and Melén, 1997).

Typically, these Gestalt-based boundary locations correspond to large register changes or either very high or very low duration ratios, that result from events with long durations or rests between events. Some performance aspects, such as breath marks or prolonged notes in Syrinx, seemed to coincide with some of the listeners' boundaries, most of which were correctly predicted by our model. Note that breath marks are not explicitly represented in the MIDI data, but they appear in the event data, just like rests, as gaps between events.

Deliège and Melén (1997) argue for a prototypical form of musical memory (and music similarity) and suggest that grouping principles are strong cues to the formation of these prototypes, but they seem to refer to Gestalt principles as being external to these learned prototypes. This view has been adopted in other melodic segmentation models where Gestalt principles are hard-coded in the form of rules (Cambouropoulos, 1998; Temperley, 2001). The MDSM, the first of our models presented in Chapter 4, did not have Gestalt rules defined explicitly, but arguably, they were implied in the definition of melodic density which depended on melodic cohesion of pitch intervals (mostly related to pitch distance) and note density (related to pitch duration).

Contrasting with this view, our second segmentation model relies on a measure of expectancy to rate boundary salience and suggest that some of these Gestalt principles might be acquired with a probabilistic learning model. It was mentioned that discontinuities in the melodic surface, such as long notes or large pitch intervals, can be represented by fairly small contexts. Although these small contexts can only represent low-level features, they benefit from greater redundancy and provide good generalisations over the training set. Feature patterns that represent discontinuity (or continuity) may become statistically salient not just because they are rare (or very probable), but because they can make strong predictions about following occurrences.

Recent experiments with listeners on structural grouping of melodies (Schaeffer et al., 2004) showed that musically experienced subjects make significantly more exceptions to the Gestalt principles when assigning phrase boundaries. In our listening study this difference between subject groups was not so apparent. On the other hand a smaller

proportion of our probabilistic model's segment predictions were not Gestalt-related, but found a correspondence in the listeners' segmentations.

An important aspect of this findings is that they show musical knowledge can have a prevailing factor in melodic segmentation. Schaeffer et al. (2004) suggest that certain musical factors, of which the experienced listener is more aware, may govern the perceived structure of melodies, and Gestalt principles can be simply overruled by these factors.

Narmour (1992) underlines the importance of these implicative processes to the formation of groupings but he argues for a separation between learning-based and innate bottom-up processes. Our results, on the other hand, suggest that grouping principles can be linked to expectations that result from acquired regularities in melodic data. This adds support to the idea (see also Pearce and Wiggins (2004)) that grouping principles can emerge from a data learning process, questioning the assumption that they result exclusively from innate universal processes of perception.

8.6 Summary

The aim of this research was to develop a computational model of music learning capable of generating perceptually pertinent segmentations for a given melody, with minimum prior knowledge given to the system. Overall, we have succeeded in building a model that can replicate the melodic segmentation intentions of human subjects, by comparing it with real listening data. But the psychological validation of a computational model requires a closer relationship between the architecture of the model and that of the human subject (Desain et al., 1998). As such, we showed how the choice of a melodic representation and the definition of some of the model's parameters were guided by research findings on music perception and cognition.

A distinctive characteristics of our approach is that the model is trained with non-annotated melodic data. This means that the model can process and generate segment boundaries for any melody provided in MIDI format, without the need for any prior processing. Data-driven approaches are attractive because the models can learn from a set of representative data and may be generalised to many different cases or even data from different domains. Such generalisations must follow the assumption that some of the principles embedded in the model are also general across different sources of data or domains. The generalised use of predictive models to learn from time series is an indication of the success of such models across domains such as linguistics, biology, and

also music. In the music domain probabilistic models have been used mostly for music generation and evaluation of musical learning, but few to model music segmentation. This work explored existing probabilistic learning techniques to model unsupervised learning of melodic information.

The next and final chapter of this dissertation highlights the contributions of this research and discusses some ideas for future work.

Chapter 9

Conclusion

This dissertation presented two data-driven memory-based models of melodic segmentation with a particular emphasis on a probabilistic approach to model musical learning. In this chapter we present a summary of this research by highlighting its main contributions. Then we identify possible developments and new directions that could be followed as future work.

9.1 Summary of Contributions

The main contributions of this research can be summarised as follows:

- Two computational models of melodic segmentation were developed exploring different aspects of musical memory. With the first of these models we showed the importance of local context for the rating of boundary strengths. Using a recency memory window and a novel concept of melodic density we re-developed an existing segmentation algorithm, the LBDM (Cambouropoulos, 1998), and improved its boundary selectivity.

  A second approach demonstrated how a model based on a simple (probabilistic) learning mechanism, can effectively learn from raw melodic data and use the acquired knowledge to predict segmentation boundaries in melodies. Experimental results with this model corroborate an initial hypothesis that boundary perception in melodies can be predicted using information theoretical principles, and support the idea that melodic expectancies play a relevant role in the perception and processing of musical data. Results also demonstrate that a very simple bottom-up learning mechanism can model and predict listening behaviour from

non-annotated melodic data. This contrasts with existing approaches which rely on hard-coded rules and domain specific data.

- Our results suggests that intra-opus regularities are paramount in the perception of boundaries and can override some aspects of long-term musical learning. It was shown that perceived boundaries were predicted predominantely based on surface discontinuities and limited length recurrent patterns. Although the small scope of the memory contexts can be seen as a limitation of the model, results are in accordance with previous evdence that surface features are paramount in the processing and recognition of music (Lamont and Dibben, 2001) .

- Approximate similarity can be embodied, to some extent, with a probabilistic learning model, but it relies heavily on the choice of melodic features. We showed that the information theoretic approach used to measure intra-opus feature salience, could be taken a step further to select the set of relevant melodic features. This highlights the potential of a careful choice of basic melodic parameters to be represented, but at the same time shows that music representation remains the great challenge in modelling musical perception.

- Our results suggest that Gestalt-based grouping principles can be linked to expectations that result from learned regularities in melodic data. This supports other research (Pearce and Wiggins, 2004) suggesting that grouping principles can emerge from a learning process, questioning the assumption that they result exclusively from innate universal processes of perception. The importance of a learning process for a segmentation task was confirmed by a proportion of the correct boundary predictions, which did not correspond to Gestalt-based principles, but were the result of acquired recurrent patterns.

- We argued for the importance of gathering real listening data for the development and evaluation of cognitively pertinent models of music listening. A listening study on melodic segmentation was carried out with two groups of subjects, with and without formal musical training. The study revealed differences between the two groups of listeners both in terms of the granularity of the segmentations and in terms of response time and anticipation. Despite these differences the study confirmed that overall, there is a significant boundary overlap between the two groups. The data produced from this listening study is made available for further research as a contribution to the study of melodic segmentation.

9.2 Future Work

Temporal aspects of melodic segmentation

- The influence of tempo in melodic segmentation, although briefly addressed during the development of the MDSM in Chapter 4, has not been accounted for in our probabilistic model. How does tempo affect segmentation and memorisation? Dramatic changes in tempo might trigger different types of phenomena. For example, we speculate that fast tempos might influence the perception of groups of notes, possibly changing the way patterns are perceived and memorised, whereas slow tempos might favour the perception of surface discontinuities. We suggest that the influence of tempo in segmentation could be the subject of an empirical study. Similarly, other real-time aspects of a performance, such as dynamics and articulation, could be addressed in further studies, to investigate how different interpretations of the same piece are reflected in the listeners' segmentation, or even how expressive interpretations of a piece compare to a dead-pan version of the same piece, in terms of segmentation.

- Palmer and Krumhansl (1987) showed that the frequency of musical events can be used to induce musical meter. It would be useful to examine the relative contributions of different kinds of phenomenal accents to the perception of meter and compare them with melodic and rhythmic salient features, as predicted by our probabilistic model. Conversely the inclusion of meter information as a complement to the data acquisition process, could be used to favour the memorisation of features in key points of a melody, for example by increasing the statistical weight of melodic sequences that occur where metrical accents are stronger. This idea would follow from recent studies suggesting that infants use metrical structure to bootstrap their knowledge acquisition in music learning (Hannon and Johnson, 2005).

Evaluation and validation

- The selection of boundary predictions was achieved using a threshold which was a function of the standard deviation of the entropy profiles. This was meant as a first approach to select boundary predictions. Larger pieces are likely to include distinct sections, with varying feature informativeness, as is it the case of Syrinx and the excerpt of Sonata K333. In these cases, a fixed threshold may not be adequate

to evaluate boundary salience in order to accommodate these melodic changes. We suggest that a dynamic selection function could be investigated. In parts of a melody where there is a lot of change, the selection of boundaries could be tightened and conversely, in parts where there are less salient features, the model could slacken the selection criterion.

- Although the aim of the present study was to compare the segment boundary predictions of the model with segmentation data obtained with listeners, it would be interesting to carry out a systematic comparison with other segmentation models such as the Grouper (Temperley, 2001) or the LBDM (Cambouropoulos, 1998). Baseline models could also be used to evaluate the quality of the predictions of our model in comparison with very simple models. Baseline models rely on informed guessing, and would predict boundaries in locations associated with certain features known to be frequently associated with the perception of segment boundaries (e.g. long notes, rests or large pitch intervals). This analysis would show if the additional complexity of our predictive model translates into higher accuracy, when compared with a chosen baseline model.

Tools for empirical studies

- In the listening study presented in Chapter 5 we mentioned the absence of a revision procedure to allow listeners to change their segmentations. The solution to this problem is non-trivial and needs further research. If one wants to include non-musically trained subjects in the experiments it is difficult to introduce any kind of notation-based visual aid, for revision purposes. Graphical representations of the melodies made available to listeners during the auditions could also lead the participants to base their segmentation on graphical, rather than auditory features. Other solutions based only on the manipulation of the sound source, like allowing "rewind" or "seek" facilities during the experiment, interfere with the idea of establishing a realistic listening experiment.

More on probabilistic learning models

- Probabilistic models are a good framework to explore different statistical measurements, to analyse the target data. For example, mutual information could be used to look at the dependencies and interaction between different melodic features.

With reference to the multiple viewpoints approach of (Conklin and Witten, 1995), different methods could be investigated to automatically select a combination of different melodic features or to evaluate different melodic representations.

- Exploratory experiments with our mixed-memory model, revealed that when very abstract melodic features such as pitch contour or duration contours were used, isolated higher order components appear as a result of the training of the model. The order of these components, seem to have some relation with some characteristics of the melodies such as the length of recurring motifs. This phenomena requires further analysis and experimentation.

Appendix A

Event lists of melodies used in the listening study

This appendix contains the event lists of the melodies used in the listening study. Events are numbered sequentially as they appear in the MIDI files. *Pitch* is denoted as MIDI note codes and event onset time (*onset*) and event duration (*dur*) are expressed in milliseconds.

id	pitch	onset	dur
1	82	3998	1000
2	81	4998	142
3	83	5141	115
4	80	5256	987
5	79	6256	167
6	81	6423	166
7	78	6589	273
8	77	6862	322
9	76	7184	388
10	73	7577	468
11	82	8048	1000
12	84	9048	148
13	83	9197	120
14	82	9318	3444
15	82	14126	1125
16	81	15251	168
17	83	15420	129
18	80	15552	1125
19	79	16677	166
20	81	16844	125
21	78	16969	375
22	77	17344	375
23	76	17719	375
24	73	18094	375
25	70	18469	1125
26	66	19594	187
27	67	19781	188
28	70	19969	1000
29	66	20969	250
30	67	21219	250
31	72	21469	250
32	65	21719	250
33	67	21969	250
34	73	22219	125
35	77	22344	125
36	76	22469	250
37	73	22719	250
38	70	22969	1125
39	66	24094	187
40	67	24281	188
41	71	24469	1566
42	68	26135	163
43	71	26299	160
44	73	26461	157
45	76	26618	154
46	80	26772	151
47	80	27596	223
48	83	27820	3582
49	85	31402	895
50	87	32297	2040
51	82	34348	1114
52	70	35950	937
53	69	36887	142
54	71	37029	113
55	68	37143	900
56	67	38043	136
57	69	38180	110
58	66	38290	311
59	65	38603	312
60	64	38915	313
61	61	39228	312
62	70	39540	918
63	69	40478	137
64	71	40617	107
65	68	40725	900
66	67	41625	134
67	69	41759	104
68	66	41863	295
69	68	42158	294
70	70	42452	294
71	73	42746	294
72	75	43040	196
73	78	43236	98
74	77	43334	294
75	75	43628	294
76	73	43922	294
77	70	44216	1323
78	68	45697	264
79	61	45961	255
80	63	46216	294
81	66	46569	138
82	68	46707	147
83	70	46863	147
84	73	47040	127
85	75	47167	89
86	78	47256	49
87	77	47344	147
88	75	47491	108
89	73	47599	147
90	70	47746	1764
91	73	49510	196
92	75	49706	196
93	78	49902	196
94	81	50098	735
95	80	50833	147
96	79	50980	147
97	78	51127	147
98	75	51274	735
99	74	52010	147
100	73	52157	147
101	72	52304	147
102	69	52451	735
103	68	53186	147
104	67	53333	147
105	66	53480	147
106	63	53627	294
107	66	53922	98
108	65	54020	392
109	64	54412	392
110	63	54804	980
111	65	55980	589
112	64	56569	588
113	63	57157	294
114	66	57451	98
115	65	57549	392
116	64	57941	392
117	63	58333	294
118	66	58627	98
119	65	58725	393
120	64	59118	392
121	63	59510	294
122	65	59804	98
123	64	59902	392
124	63	60294	392
125	62	60686	254
126	65	60942	77
127	64	61019	309
128	61	61328	345
129	60	61673	510
130	61	62185	451
131	66	62637	590
132	69	63227	676
133	62	63912	502
134	66	64420	156
135	65	64576	527
136	62	65104	391
137	61	65496	1556
138	73	67060	818
139	62	68202	319
140	65	68521	106
141	64	68627	426
142	61	69053	425
143	60	69478	540
144	61	70026	486
145	66	70514	415
146	69	70930	374
147	62	71305	294
148	66	71600	98
149	65	71698	392
150	62	72090	391
151	61	72482	1291
152	75	73783	2003
153	73	75816	200
154	71	76016	197
155	70	76213	192
156	68	76405	1222
157	63	77722	181
158	66	77904	139
159	68	78045	118
160	70	78164	111
161	73	78275	97
162	75	78373	256
163	73	78631	85
164	75	78718	86
165	77	78804	86
166	75	78890	127
167	73	79018	129
168	71	79148	149
169	70	79298	1023
170	71	80384	120
171	73	80506	119
172	74	80626	118
173	73	80745	116
174	71	80861	113
175	70	80975	112
176	71	81087	113
177	74	81200	112
178	76	81312	113
179	74	81425	113
180	71	81538	113
181	70	81651	83
182	69	81735	78
183	68	81813	77
184	66	81891	76
185	62	81968	76
186	66	82045	75
187	68	82120	75
188	69	82195	74
189	70	82269	458
190	75	82848	69
191	73	82918	70
192	75	82989	70
193	73	83060	69
194	75	83130	70
195	73	83200	71
196	75	83272	70
197	73	83343	56
198	75	83399	69
199	73	83482	60
200	75	83542	68
201	73	83620	70
202	75	83690	71
203	73	83761	70
204	75	83831	69
205	73	83900	71
206	75	83971	71
207	73	84042	69
208	75	84111	69
209	73	84180	71
210	75	84251	69
211	73	84320	68
212	75	84388	70
213	73	84458	71
214	75	84529	70
215	73	84599	71
216	75	84670	69
217	73	84739	68
218	75	84807	70
219	73	84877	71
220	75	84948	71
221	73	85019	69
222	70	85088	1090
223	78	86278	72
224	77	86350	72
225	78	86422	73
226	77	86495	73
227	78	86568	74
228	77	86642	75
229	78	86717	74
230	77	86791	75
231	78	86866	76
232	77	86942	77
233	78	87019	78
234	77	87097	77
235	78	87174	77
236	77	87251	77
237	78	87328	80
238	77	87408	82
239	78	87490	82
240	77	87572	84
241	78	87656	84
242	77	87740	85
243	78	87825	88
244	77	87913	89
245	78	88002	90
246	77	88092	94
247	70	88186	538
248	77	88798	93
249	78	88892	84
250	80	88977	81
251	82	89059	3497
252	82	94342	1706
253	81	96049	144
254	83	96193	110
255	80	96303	957
256	79	97261	143
257	81	97404	110
258	78	97514	306
259	77	97826	315
260	76	98145	391
261	73	98541	332
262	82	99073	1415
263	85	100488	123
264	88	100611	160
265	86	100771	283
266	85	101054	283
267	82	101337	2072
268	81	103418	163
269	83	103584	165
270	80	103751	1033
271	79	104793	179
272	81	104973	182
273	78	105161	371
274	77	105536	382
275	76	105918	393
276	73	106311	269
277	70	106716	277
278	69	107002	275
279	71	107287	276
280	68	107573	276
281	67	107859	275
282	69	108144	276
283	66	108430	416
284	65	108859	418
285	64	109287	419
286	61	109716	277
287	70	110144	276
288	69	110430	276
289	71	110716	276
290	68	111002	275
291	67	111287	276
292	69	111573	276
293	66	111859	422
294	65	112294	447
295	64	112741	476
296	61	113226	1776
297	61	115152	290
298	66	115450	174
299	65	115629	465
300	64	116105	466
301	61	116581	3255
302	69	119914	465
303	67	120390	466
304	61	120866	2674
305	71	123724	3579
306	69	127319	488
307	67	127814	484
308	65	128298	496
309	63	128794	539
310	63	129392	697
311	61	130106	4186

Figure A.1: Event list for Syrinx

id	pitch	onset	dur
1	69	832	416
2	69	1248	417
3	74	1665	832
4	73	2497	209
5	74	2706	208
6	76	2914	208
7	78	3122	209
8	79	3331	208
9	76	3539	208
10	79	3747	209
11	76	3956	208
12	74	4164	208
13	73	4372	209
14	71	4581	208
15	69	4789	208
16	81	4997	209
17	78	5206	208
18	81	5414	208
19	78	5622	209
20	83	5831	208
21	79	6039	208
22	83	6247	209
23	79	6456	208
24	78	6664	417
25	76	7081	416
26	69	7497	417
27	69	7914	417
28	74	8331	832
29	78	9163	209
30	76	9372	208
31	74	9580	208
32	73	9788	209
33	71	9997	208
34	74	10205	208
35	71	10413	209
36	74	10622	208
37	83	10830	208
38	81	11038	209
39	80	11247	208
40	78	11455	208
41	76	11663	209
42	81	11872	208
43	78	12080	208
44	74	12288	209
45	73	12497	208
46	76	12705	208
47	74	12913	209
48	71	13122	208
49	71	13330	417
50	69	13747	416
51	69	14163	417
52	69	14580	417
53	74	14997	832
54	73	15829	209
55	74	16038	208
56	76	16246	208
57	78	16454	209
58	79	16663	208
59	76	16871	208
60	79	17079	209
61	76	17288	208
62	74	17496	208
63	73	17704	209
64	71	17913	208
65	69	18121	208
66	81	18329	209
67	78	18538	208
68	81	18746	208
69	78	18954	209
70	83	19163	208
71	79	19371	208
72	83	19579	209
73	79	19788	208
74	78	19996	417
75	76	20413	416
76	69	20829	417
77	69	21246	417
78	74	21663	832
79	78	22495	209
80	76	22704	208
81	74	22912	208
82	73	23120	209
83	71	23329	208
84	74	23537	208
85	71	23745	209
86	74	23954	208
87	83	24162	208
88	81	24370	209
89	80	24579	208
90	78	24787	208
91	76	24995	209
92	81	25204	208
93	78	25412	208
94	74	25620	209
95	73	25829	208
96	76	26037	208
97	74	26245	209
98	71	26454	208
99	71	26662	417
100	69	27079	416
101	81	27495	209
102	80	27704	208
103	81	27912	208
104	80	28120	209
105	81	28329	208
106	83	28537	208
107	80	28745	208
108	76	28953	209
109	79	29162	208
110	81	29370	208
111	78	29578	209
112	74	29787	208
113	86	29995	208
114	85	30203	209
115	83	30412	208
116	82	30620	208
117	83	30828	209
118	81	31037	208
119	79	31245	208
120	78	31453	209
121	79	31662	416
122	78	32078	417
123	76	32495	417
124	74	32912	416
125	78	33328	833
126	76	34161	208
127	69	35827	417
128	69	36244	416
129	74	36660	833
130	73	37493	208
131	74	37701	208
132	76	37909	209
133	78	38118	208
134	79	38326	208
135	76	38534	209
136	79	38743	208
137	76	38951	208
138	74	39159	209
139	73	39368	208
140	71	39576	208
141	69	39784	209
142	81	39993	208
143	86	40201	208
144	83	40409	209
145	79	40618	208
146	78	40826	208
147	81	41034	209
148	79	41243	208
149	76	41451	208
150	76	41659	417
151	74	42076	417
152	81	42493	208
153	80	42701	208
154	81	42909	209
155	80	43118	208
156	81	43326	208
157	83	43534	209
158	80	43743	208
159	76	43951	208
160	79	44159	209
161	81	44368	208
162	78	44576	208
163	74	44784	209
164	86	44993	208
165	85	45201	208
166	83	45409	209
167	82	45618	208
168	83	45826	208
169	81	46034	209
170	79	46243	208
171	78	46451	208
172	79	46659	417
173	78	47076	417
174	76	47493	416
175	74	47909	417
176	78	48326	833
177	76	49159	208
178	69	50825	416
179	69	51241	417
180	74	51658	832
181	73	52490	209
182	74	52699	208
183	76	52907	208
184	78	53115	209
185	79	53324	208
186	76	53532	208
187	79	53740	209
188	76	53949	208
189	74	54157	208
190	73	54365	209
191	71	54574	208
192	69	54782	208
193	81	54990	209
194	86	55199	208
195	83	55407	208
196	79	55615	209
197	78	55824	208
198	81	56032	208
199	79	56240	209
200	76	56449	208
201	76	56657	416
202	74	57074	417

Figure A.2: Event list for K284 (extract)

id	pitch	onset	dur	id	pitch	onset	dur	id	pitch	onset	dur
1	77	1499	124	71	86	17247	125	141	77	32994	125
2	75	1623	125	72	87	17372	125	142	84	33119	125
3	74	1748	125	73	89	17497	250	143	83	33244	125
4	72	1873	125	74	89	17747	250	144	84	33369	125
5	72	1998	250	75	89	17997	250	145	86	33494	125
6	70	2248	250	76	87	18247	500	146	84	33619	125
7	70	2498	750	77	86	18747	250	147	82	33744	125
8	74	3248	250	78	86	18997	250	148	81	33869	125
9	79	3498	250	79	84	19247	250	149	81	33994	250
10	74	3748	250	80	82	19497	250	150	79	34244	250
11	77	3998	500	81	81	19747	250	151	79	34744	250
12	75	4498	749	82	82	19997	500	152	79	34994	250
13	77	5247	125	83	70	20497	500	153	77	35244	250
14	79	5372	125	84	67	21496	125	154	77	35744	250
15	77	5497	125	85	65	21621	125	155	76	35994	125
16	75	5622	125	86	63	21746	125	156	77	36119	125
17	74	5747	125	87	62	21871	125	157	76	36244	125
18	72	5872	125	88	60	21996	250	158	72	36369	125
19	70	5997	500	89	58	22246	250	159	77	36494	125
20	69	6497	750	90	58	22496	749	160	79	36619	125
21	69	7247	250	91	70	23245	250	161	77	36744	125
22	70	7497	250	92	69	23495	250	162	72	36869	125
23	72	7747	250	93	67	23745	250	163	79	36994	125
24	72	7997	500	94	65	23995	500	164	81	37119	125
25	74	8497	500	95	64	24495	625	165	79	37244	125
26	77	9247	125	96	67	25120	125	166	72	37369	125
27	76	9372	125	97	65	25245	125	167	81	37494	125
28	79	9497	125	98	69	25370	125	168	82	37619	125
29	77	9622	125	99	67	25495	125	169	81	37744	125
30	75	9747	125	100	70	25620	125	170	72	37869	125
31	74	9872	125	101	69	25745	125	171	82	37994	250
32	74	9997	250	102	72	25870	125	172	82	38244	250
33	72	10247	500	103	70	25995	125	173	79	38744	250
34	70	10747	250	104	67	26120	125	174	82	38994	250
35	74	10997	250	105	76	26245	125	175	81	39244	250
36	72	11247	500	106	72	26370	125	176	79	39494	250
37	69	11747	250	107	79	26495	125	177	77	39744	250
38	70	11997	125	108	76	26620	125	178	76	39994	125
39	72	12122	125	109	82	26745	125	179	72	40119	125
40	70	12247	125	110	81	26870	125	180	83	40244	125
41	69	12372	125	111	82	26995	125	181	84	40369	125
42	70	12497	125	112	81	27120	125	182	77	40494	125
43	72	12622	125	113	79	27245	125	183	72	40619	125
44	74	12747	125	114	77	27370	125	184	83	40744	125
45	75	12872	125	115	76	27495	125	185	84	40869	125
46	76	12997	125	116	74	27620	125	186	79	40994	125
47	77	13122	125	117	72	27745	125	187	72	41119	125
48	76	13247	125	118	70	27870	125	188	83	41244	125
49	77	13372	125	119	70	27995	125	189	84	41369	125
50	79	13497	125	120	69	28120	125	190	81	41494	125
51	77	13622	125	121	74	28245	125	191	72	41619	125
52	75	13747	125	122	72	28370	125	192	83	41744	125
53	74	13872	125	123	72	28495	500	193	84	41869	125
54	74	13997	250	124	86	29494	125	194	82	41994	250
55	72	14247	250	125	84	29619	125	195	82	42244	250
56	72	14497	250	126	82	29744	125	196	79	42744	250
57	70	14747	250	127	81	29869	125	197	82	42994	250
58	74	14997	250	128	79	29994	750	198	81	43244	250
59	72	15247	250	129	76	30744	250	199	79	43494	250
60	72	15497	250	130	77	30994	125	200	77	43744	250
61	69	15747	250	131	84	31119	125	201	76	43994	125
62	70	15997	250	132	83	31244	125	202	84	44119	125
63	72	16247	125	133	84	31369	125	203	79	44244	125
64	74	16372	125	134	86	31494	125	204	76	44369	125
65	75	16497	125	135	84	31619	125	205	72	44494	125
66	77	16622	125	136	82	31744	125	206	67	44619	125
67	79	16747	125	137	81	31869	125	207	64	44744	125
68	81	16872	125	138	79	31994	250	208	60	44869	125
69	82	16997	125	139	79	32244	500	209	48	44994	500
70	84	17122	125	140	76	32744	250				

Figure A.3: Event list for K333 (extract)

id	pitch	onset	dur
1	63	0	400
2	67	399	400
3	68	799	400
4	70	1199	600
5	72	1799	200
6	70	1999	400
7	68	2399	400
8	67	2799	400
9	65	3199	400
10	67	3599	800
11	63	4399	400
12	63	4799	400
13	67	5199	400
14	68	5599	400
15	70	5999	600
16	72	6599	200
17	70	6799	400
18	68	7199	400
19	67	7599	400
20	65	7999	400
21	67	8399	800
22	63	9199	400
23	65	9599	400
24	65	9999	400
25	67	10399	400
26	68	10799	600
27	70	11399	200
28	68	11599	400
29	67	11999	400
30	67	12399	400
31	68	12799	400
32	70	13199	600
33	72	13799	200
34	70	13999	400
35	68	14399	400
36	67	14799	400
37	65	15199	400
38	65	15599	800
39	63	16399	400

Figure A.4: Event list for folk-song E0547

id	pitch	onset	dur
1	72	1199	450
2	70	1649	150
3	67	1799	300
4	63	2099	300
5	63	2399	300
6	63	2699	300
7	72	2999	300
8	70	3299	300
9	68	3599	300
10	65	3899	300
11	65	4199	300
12	65	4499	300
13	70	4799	450
14	70	5249	150
15	74	5399	300
16	74	5699	300
17	74	5999	300
18	72	6299	300
19	70	6599	300
20	68	6899	300
21	67	7199	300
22	63	7499	300
23	63	7799	300
24	63	8099	300
25	70	8399	300
26	70	8699	300
27	67	8999	300
28	63	9299	300
29	63	9599	300
30	63	9899	300
31	75	10199	300
32	75	10499	300
33	75	10799	300
34	65	11099	300
35	65	11399	300
36	65	11699	300
37	75	11999	300
38	75	12299	300
39	75	12599	300
40	74	12899	300
41	72	13199	300
42	70	13499	300
43	68	13799	300
44	62	14099	300
45	63	14399	1200

Figure A.5: Event list for folk-song F0927

id	pitch	onset	dur
1	62	899	300
2	62	1199	300
3	69	1499	300
4	69	1799	300
5	69	2099	300
6	71	2399	300
7	69	2699	300
8	67	2999	600
9	69	3599	200
10	71	3799	200
11	73	3999	200
12	74	4199	600
13	69	4799	600
14	62	5699	300
15	62	5999	300
16	69	6299	300
17	69	6599	300
18	69	6899	300
19	71	7199	300
20	69	7499	300
21	67	7799	600
22	69	8399	200
23	71	8599	200
24	73	8799	200
25	74	8999	600
26	69	9599	600
27	69	10499	150
28	69	10649	150
29	74	10799	200
30	74	10999	200
31	74	11199	200
32	72	11399	300
33	69	11699	300
34	72	11999	300
35	72	12299	300
36	65	12599	600
37	67	13199	200
38	67	13399	200
39	67	13599	200
40	69	13799	200
41	69	13999	200
42	69	14199	200
43	70	14399	300
44	67	14699	300
45	69	14999	600
46	67	15599	300
47	64	15899	300
48	65	16199	300
49	62	16499	300
50	64	16799	300
51	64	17099	300
52	62	17399	600
53	69	17999	200
54	69	18199	200
55	69	18399	200
56	74	18599	600
57	69	19199	600

Figure A.6: Event list for folk-song Q0034

Appendix B

Kernel Density Estimation

Kernel density estimation (KDE) provides a way of approximating the probability density function of a random variable. The following overview of KDE is mostly based on (Silverman, 1986) to which the reader is referred for additional details.

Assuming that we have n observations $x_1, x_2, ..., x_n$ from a random variable X, the kernel density estimator $\hat{f}_h(x)$ is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) \tag{B.1}$$

where K denotes a kernel function and h denotes the window width, also often referred to as the bandwidth or smoothing parameter.

The kernel estimator is thus a sum of 'bumps' placed at the points of the observations. The kernel function $K(u)$ determines the shape of these 'bumps' and the bandwidth their width. The bandwidth h affects the smoothness of the distribution: the larger the value of h the smoother the resulting distributions and vice versa.

Usually, the kernel function K will be a symmetric probability density function which satisfies the condition

$$\int_{-\infty}^{+\infty} K(x)dx = 1 \tag{B.2}$$

Table B.1 shows a few possible kernel functions.

KDE constitutes an alternative method to histograms, with the advantage that it does not depend on the choice of an origin. However there is still the need to define the window width, in many ways comparable to the problematic choice of bin width in histograms. The choice of h is important, since there is a compromise between the smooth-

Kernel	$K(X)$
Epanechnikov	$\begin{cases} \frac{\frac{3}{4}(1-\frac{1}{5}x^2)}{\sqrt{5}} & \text{if } \lvert x\rvert < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$
Biweight	$\begin{cases} \frac{15}{16}(1-x)^2 & \text{if } \lvert x\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$
Triangular	$\begin{cases} 1-\lvert x\rvert & \text{if } \lvert x\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
Rectangular	$\begin{cases} \frac{1}{2} & \text{if } \lvert x\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$

Table B.1: Kernel functions

ing and the resolution of the estimated distribution. Silverman (1986) suggests that for many applications a subjective choice of h may suffice, highlighting the advantages of examining several plots with different smoothing values over a single value obtained automatically. The reader is also referred to (Silverman, 1986), for a discussion on alternative methods of choosing a smoothing parameter.

Appendix C

Listening study

This appendix provides additional information about the listening study presented in Chapter 5 which includes:

- Overview of the participants in the listening study
- Copy of the instruction sheet/questionnaire handed to the participants
- Probability density graphs of listeners' segment boundaries (full result data from the listening study can be found in the CD annexed to this dissertation).

Subject ID	Age	Sex	Group (Mus. / Non-Mus)	Musical studies A	B	C	D	E	F	Instrument	Years played
1	24	M	M						X	Saxophone	19
2	28	M	M		X	X					18
3	21	M	M		X					Guitar	8
4	29	M	M			X	X	X	X	Cello	6
5	26	M	M				X	X	X	Organ	19
6	30	F	M				X	X	X		19
7	23	F	N-M							Piano	5
8	20	M	M			X				Piano	12
9	22	F	N-M	X							
10	25	F	M			X	X			Piano	18
11	23	F	M				X			Voice/Piano	15
12	24	F	M				X			Saxophone	11
13	20	M	M		X					Piano	12
14	19	M	M			X				Piano	9
15	20	M	M			X				Flute	10
16	26	M	N-M		X						
17	22	M	N-M		X						
18	23	F	N-M		X						
19	25	F	N-M	X							
20	27	M	M			X				Piano	10
21	27	F	N-M		X					Clarinet	6
22	24	M	N-M	X							
23	38	M	N-M		X						
24	31	M	N-M		X					Guitar	2
25	22	F	N-M	X							
26	26	F	M			X				Piano	15
27	23	M	N-M		X					Piano	4
28	21	M	M			X				Piano	12
29	21	F	M			X				Piano	15
30	27	M	M		X					Percussion	12
31	25	F	N-M		X					Piano	n.s
32	21	F	N-M	X	X						
33	24	F	M				X			Piano	18
34	24	M	N-M		X					Piano	n.s.
35	21	F	N-M		X						
36	30	F	M				X	X	X	Violin	17
37	25	F	M		X					Violin	9
38	21	F	N-M		X						
39	23	M	N-M	X						Piano	5
40	24	F	M		X					Piano	20
41	27	M	M							French Horn	13
42	24	M	N-M	X							
43	28	F	N-M	X							
44	26	F	N-M	X							
45	20	F	N-M		X					Guitar	1
46	20	F	N-M		X					Viola	3
47	24	M	N-M		X					Piano	3
48	26	M	M				X		X		14

A: Has never studied music
B: Studied music but only in the basic school
C: Is a student in a Music School
D: Is a music graduate
E: Is a music teacher
F: Is a professional musician

Figure C.1: Overview of the participants in the listening study

Subject no	Musician ¤	Non musician ¤	Date .

Introduction

The aim of this session is to study some of the factors that are used by listeners to divide a melodic passage in several parts or sections. We call this melodic segmentation.

During this session you will listen to several melodic samples while interacting with a computer program. The program will guide you on all necessary steps and provide you instructions as you go along. Please read and follow them carefully.

Procedure

You will be listening to 6 different melodic pieces. Try to imagine that each melody is a short story line that you have to break down in several smaller episodes (or segments).

You can indicate a division between segments, by pressing the large Segment button, while the melody is being played. This button will not be activated during the familiarisation auditions.

For every melody you will be given the opportunity to have two familiarisation auditions, followed by one segmentation practice round and then a final segmentation round. Only the final segmentation round will be recorded.

These sessions are not timed so take you time!
Take a few seconds rest, before the start of each new melody.

There is no such thing as a correct segmentation. We are looking for **your** individual perception of the segments in these melodies.

Before you start

Please fill in the information below.

If you have any questions, during the experiments, please raise your hand to attract our attention: we will try to provide all necessary clarifications. Please avoid, asking questions to other participants.

Please let us know when you re ready to start.

Thank your for participating.

Course of studies attended at present ..

Age Sex: M ¤ F ¤

Musical studies (please tick the relevant boxes):

- ¤ I have never studied music
- ¤ I studied music but only in the basic school
- ¤ I m a student in a Music School
- ¤ I m a music graduate
- ¤ I m music teacher
- ¤ I m a professional musician

¤ I have played an instrument for years . Instrument: ...

Figure C.2: Copy of instruction sheet/questionnaire handed to the participants.

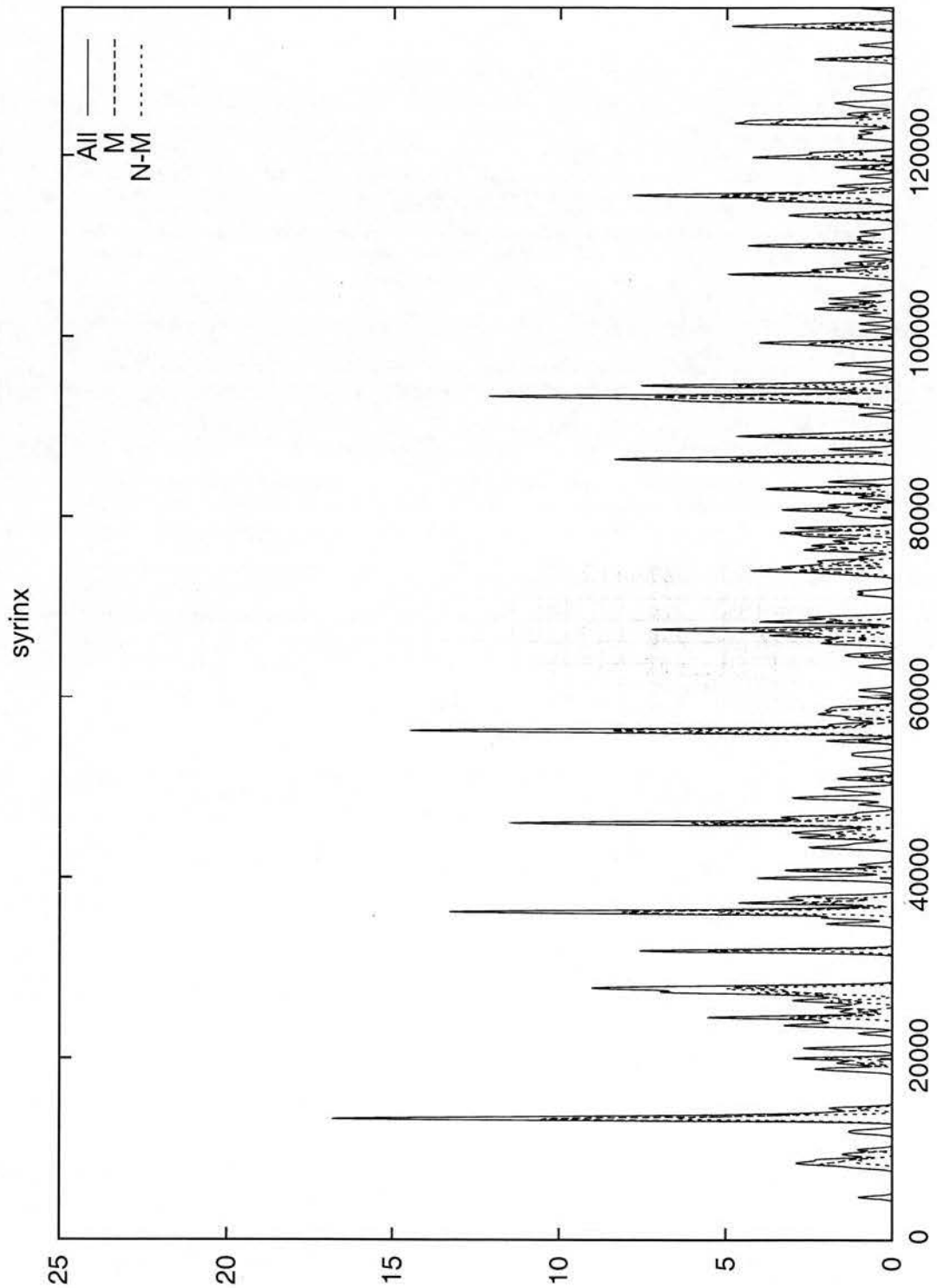


Figure C.3: Probability density of listeners' segment boundaries for Syrinx

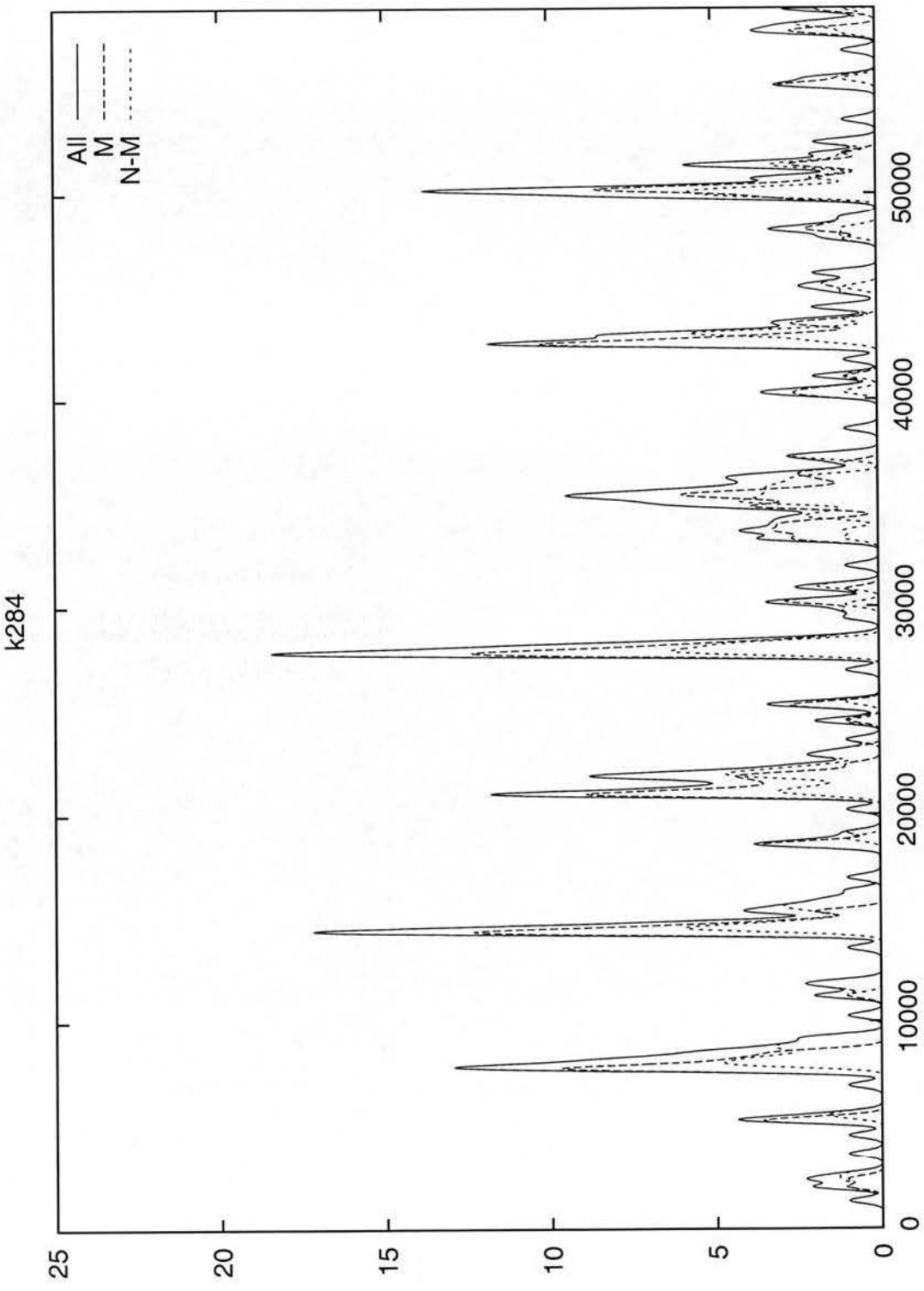


Figure C.4: Probability density of listeners' segment boundaries for K284

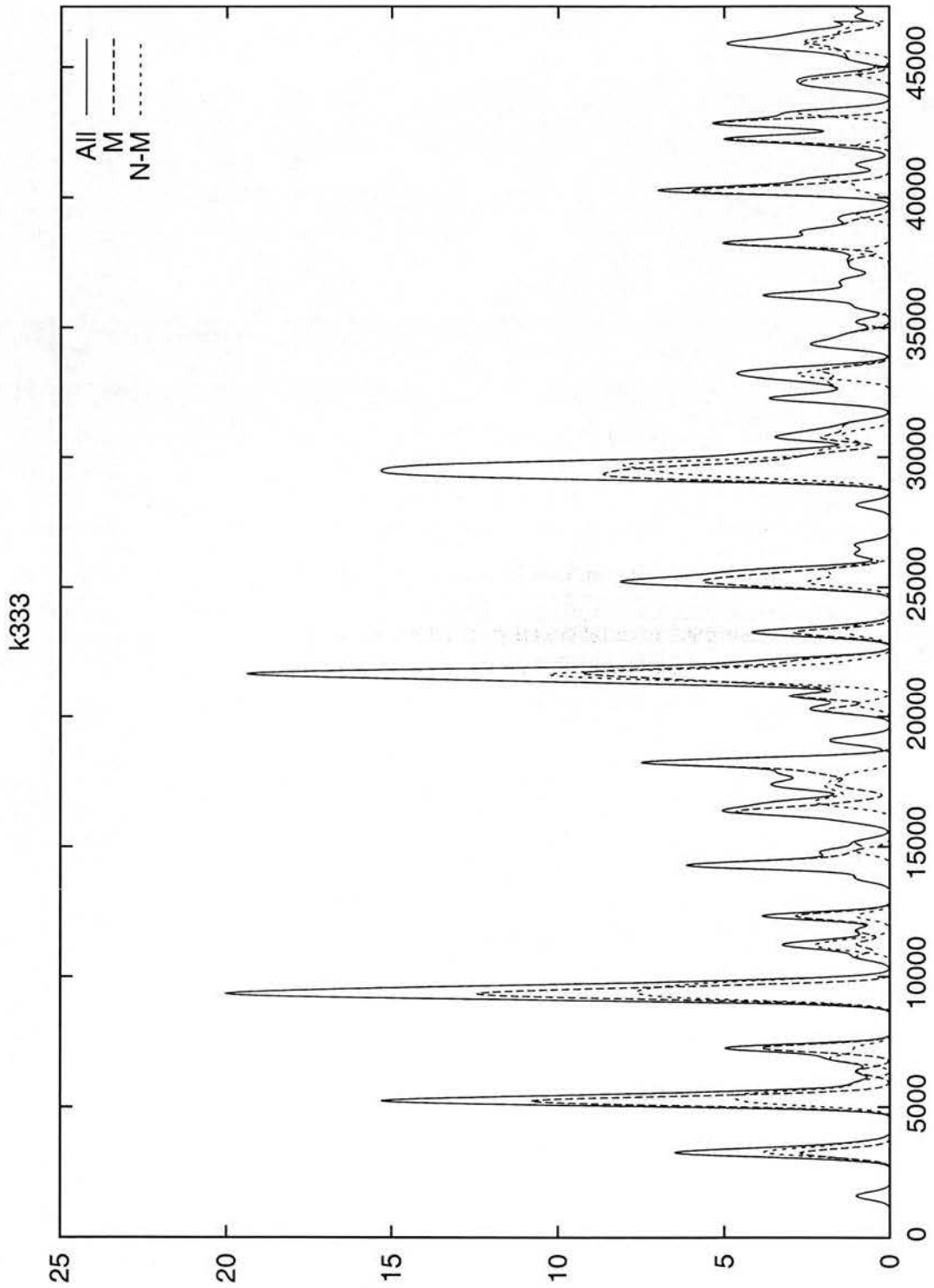


Figure C.5: Probability density of listeners' segment boundaries for K333

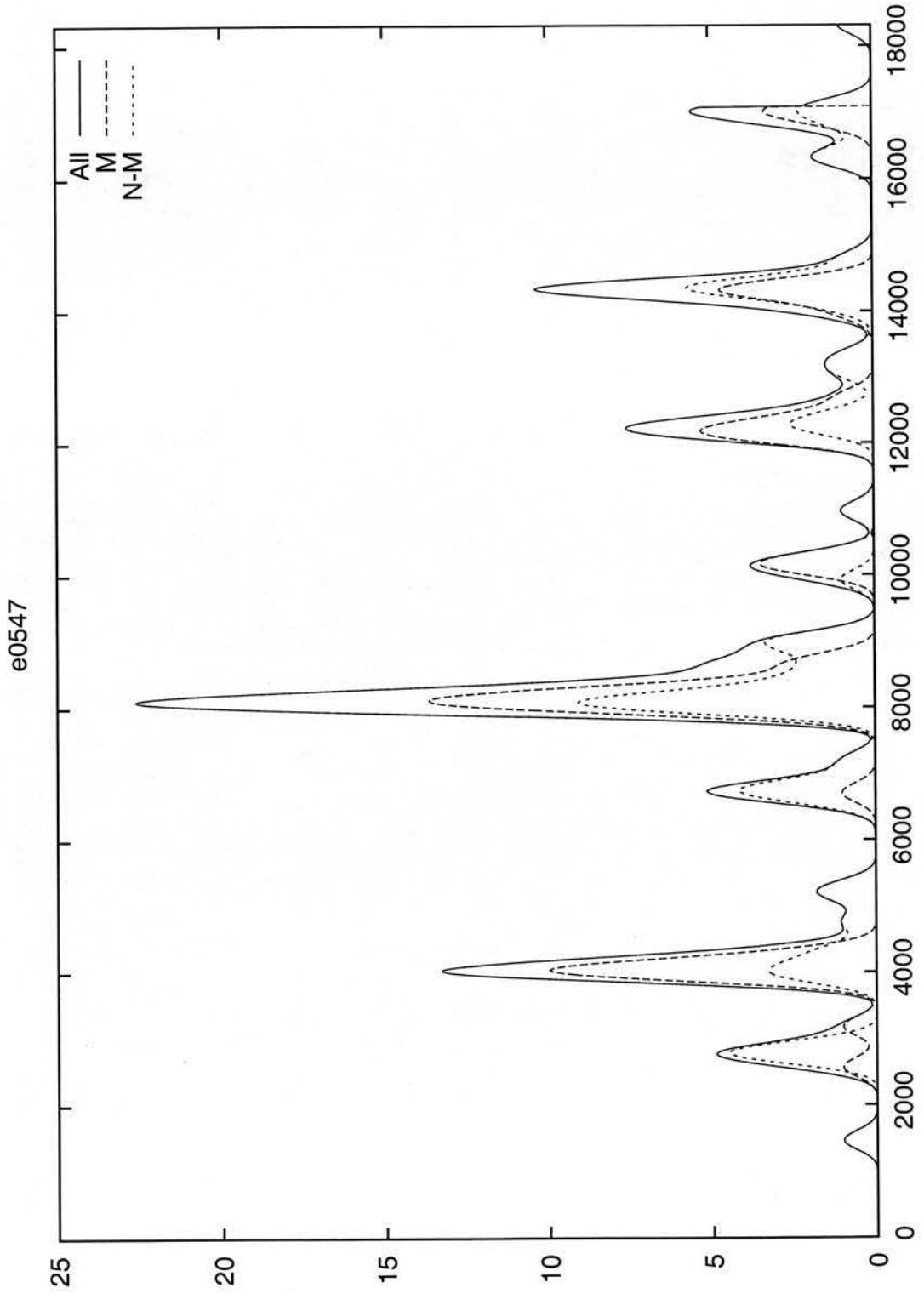


Figure C.6: Probability density of listeners' segment boundaries for E0547

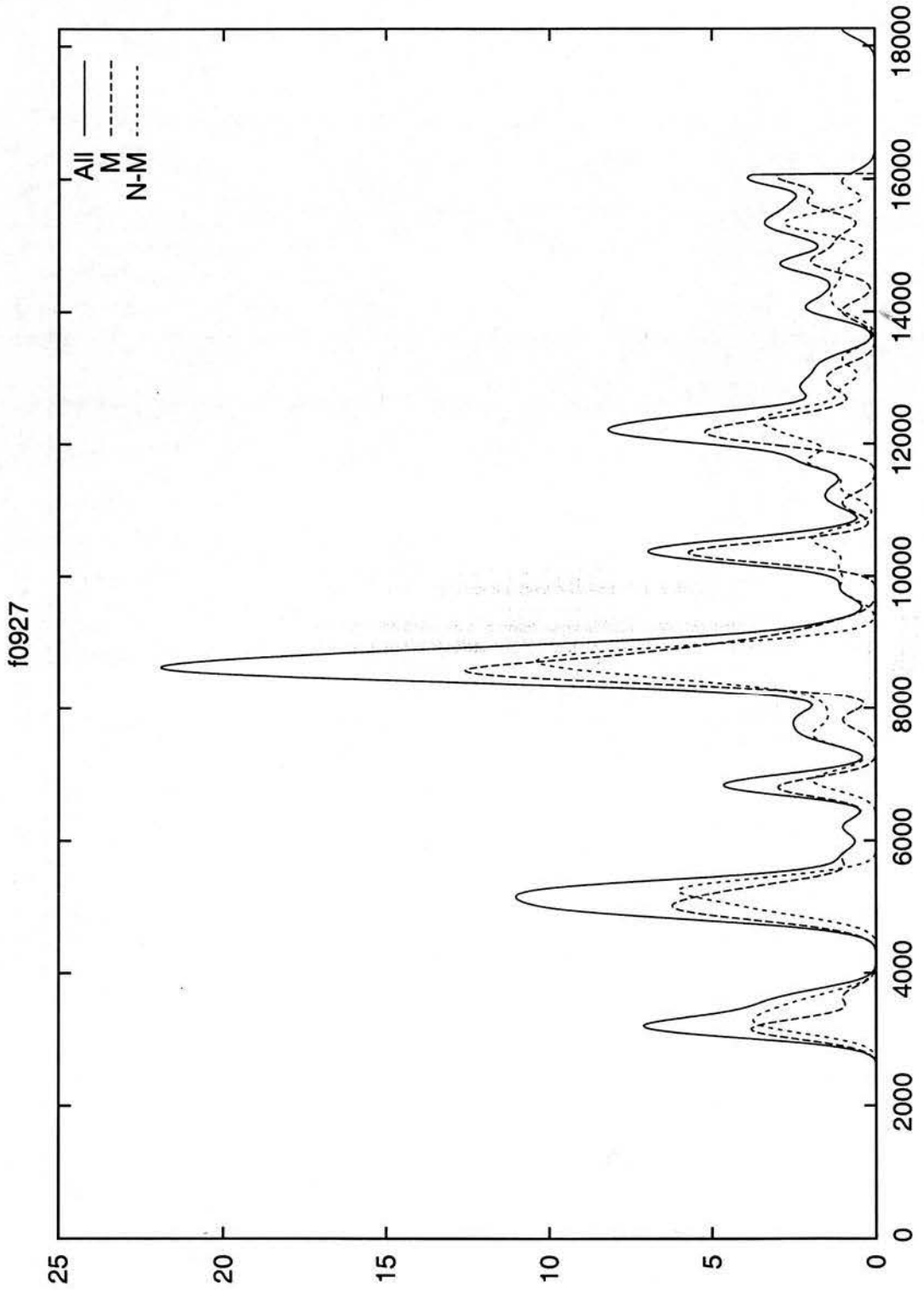


Figure C.7: Probability density of listeners' segment boundaries for F0927

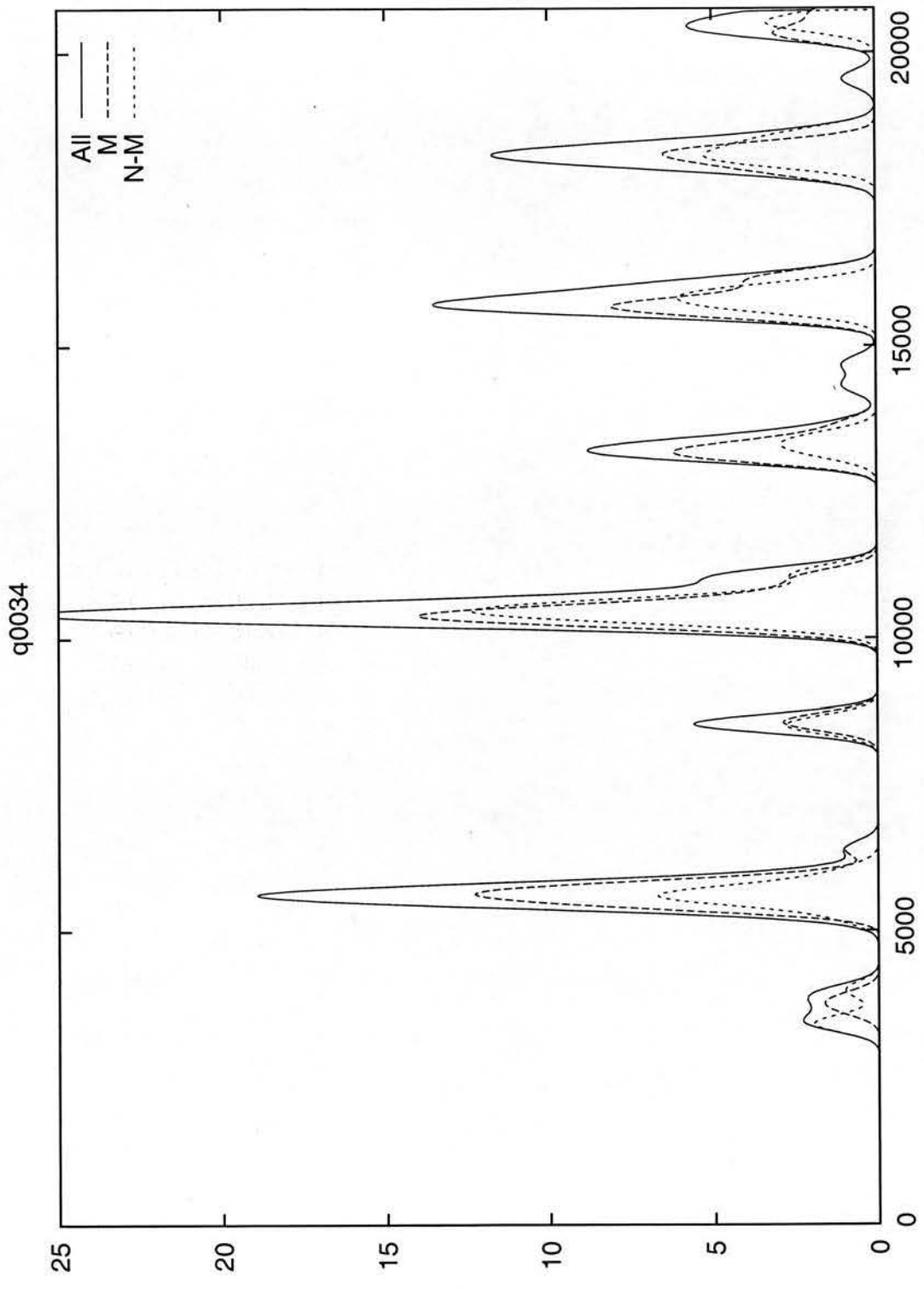


Figure C.8: Probability density of listeners' segment boundaries for Q0034

Appendix D

Experimental Results Data

This appendix includes all data charts and tables corresponding to the experimental results presented in Chapter 7

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
13360	100	14,15
24560	33	41,42
27780	54	47,48
31910	45	49,50
36200	79	52,53
46050	68	79,80
56270	86	111,112
67590	43	138,139
86410	50	224,225
93370	72	251,252
94600	45	252,253
115670	46	299,300

Table D.1: Listener boundaries selected for Syrinx. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
7850	70	26,27
14410	65	51,52
21030	64	76,77
27810	100	102,103
35390	51	126,127
42750	64	153,154
50200	75	177,178

Table D.2: Listener boundaries selected for K284. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

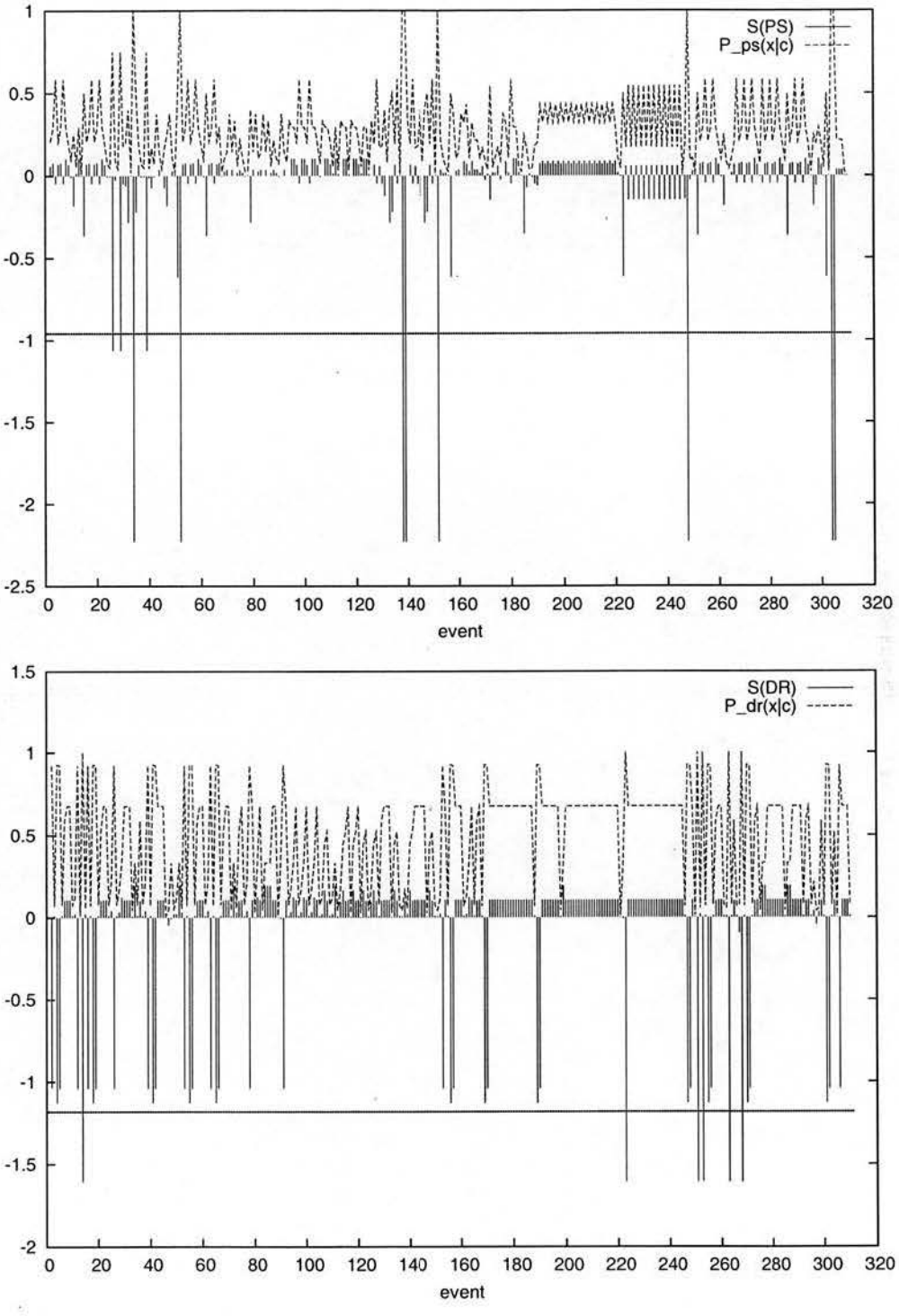


Figure D.1: Syrinx: boundary predictions $S(c)$ and successor probability $P(x|c)$ for features PS and DC. Boundary selection threshold indicated a dotted line at the bottom of the graph

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
5240	76	12,13
9350	100	26,27
18240	37	75,76
21680	97	85,86
25220	41	96,97
29470	77	123,124
40280	35	180,181

Table D.3: Listener boundaries selected for K333. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
4040	59	10,11
8110	100	20,21
12230	34	29,30
14340	46	34,35

Table D.4: Listener boundaries selected for melody E0547. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

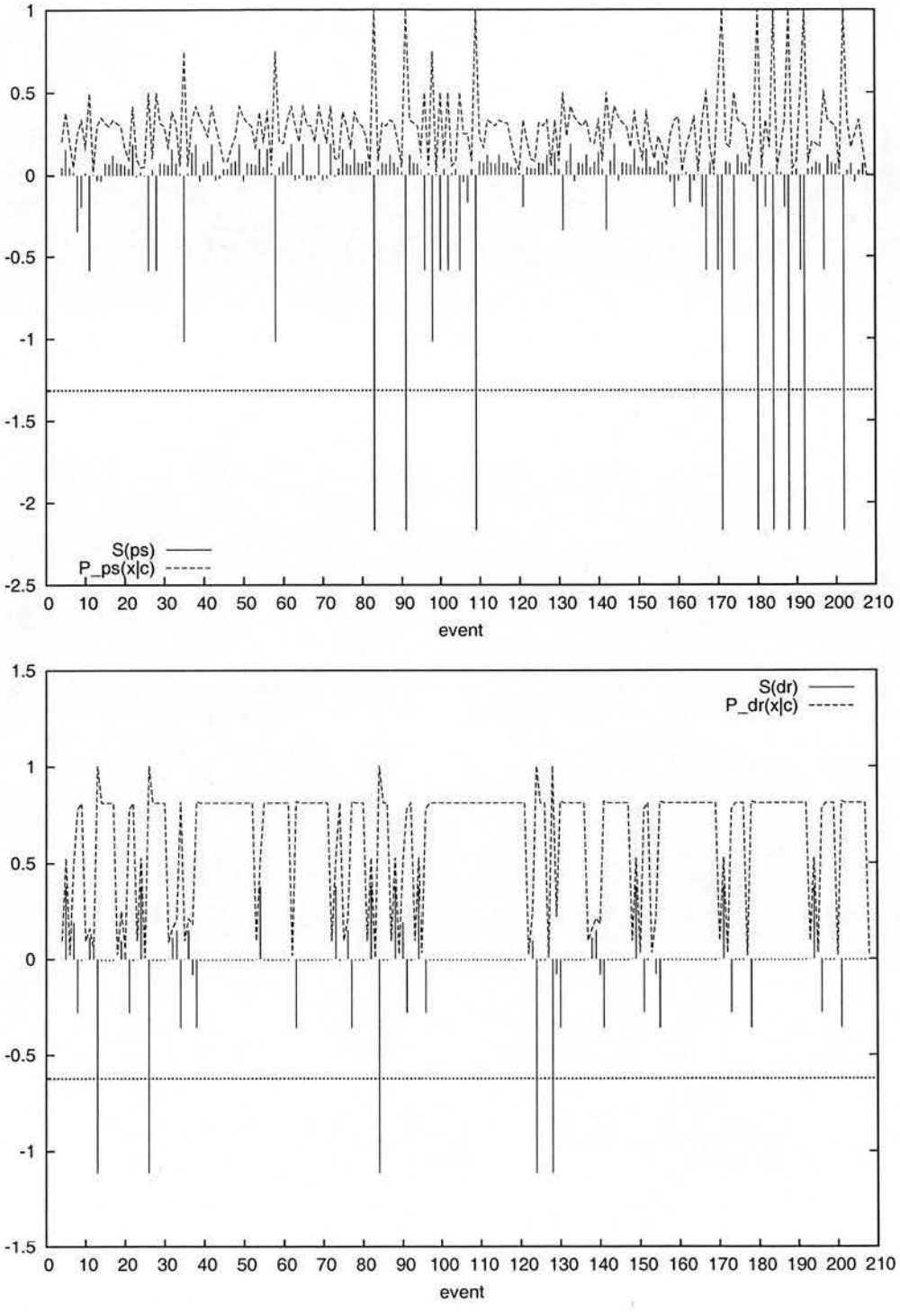


Figure D.2: Sonata K333: boundary predictions $S(C)$ and successor probability $P(X|C)$ for features PS and DC. Boundary selection threshold indicated by a dotted line at the bottom of the graph

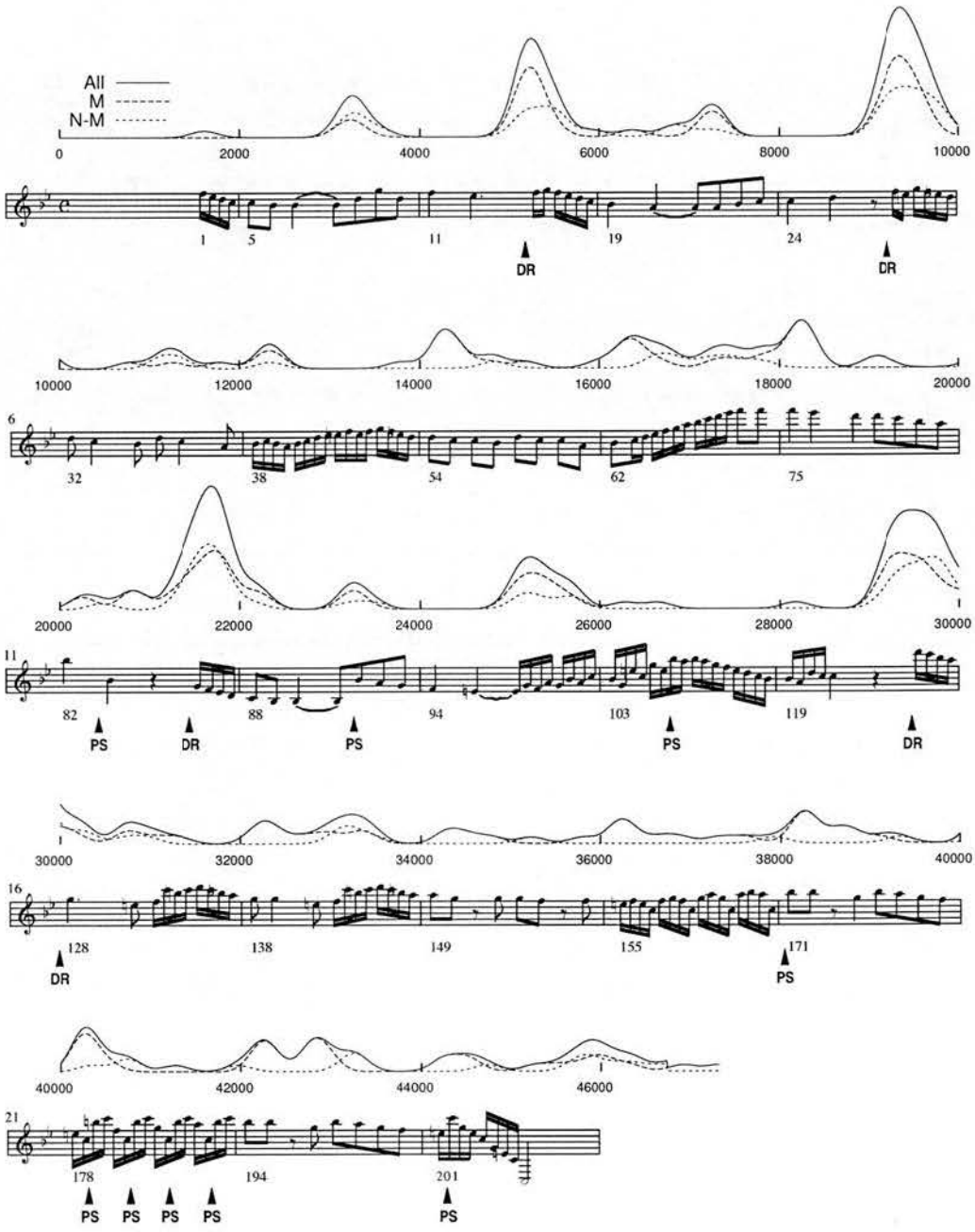


Figure D.3: Comparison between model boundary predictions and listeners' boundaries for melody K333

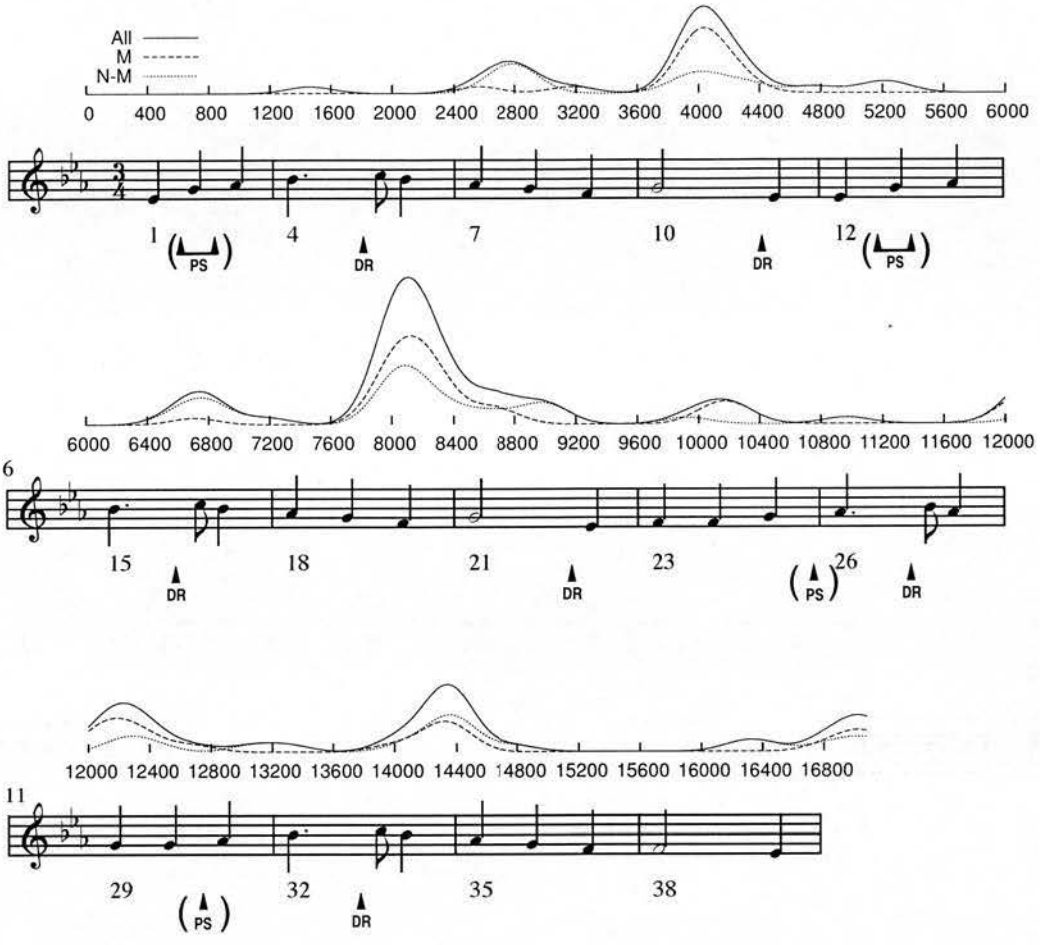


Figure D.4: Comparison between model boundary predictions and listeners' boundaries for melody E0547

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
5150	50	13,14
8630	100	25,26
12220	37	37,38

Table D.5: Listener boundaries selected for melody F0927. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

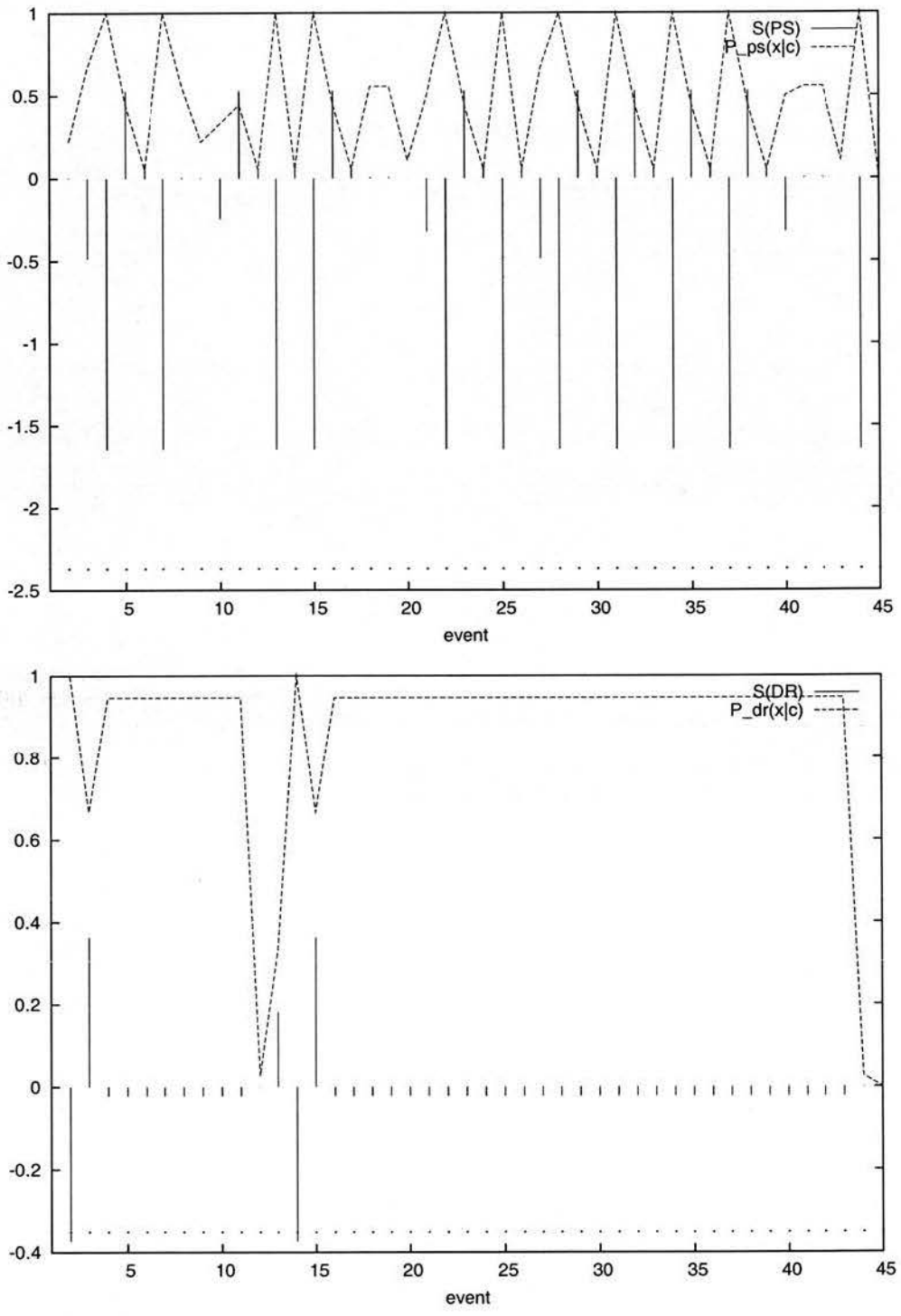


Figure D.5: F0927: boundary predictions $S(C)$ and successor probability $P(X|C)$ for features PS and DC. Boundary selection threshold indicated by a dotted line at the bottom of the graph

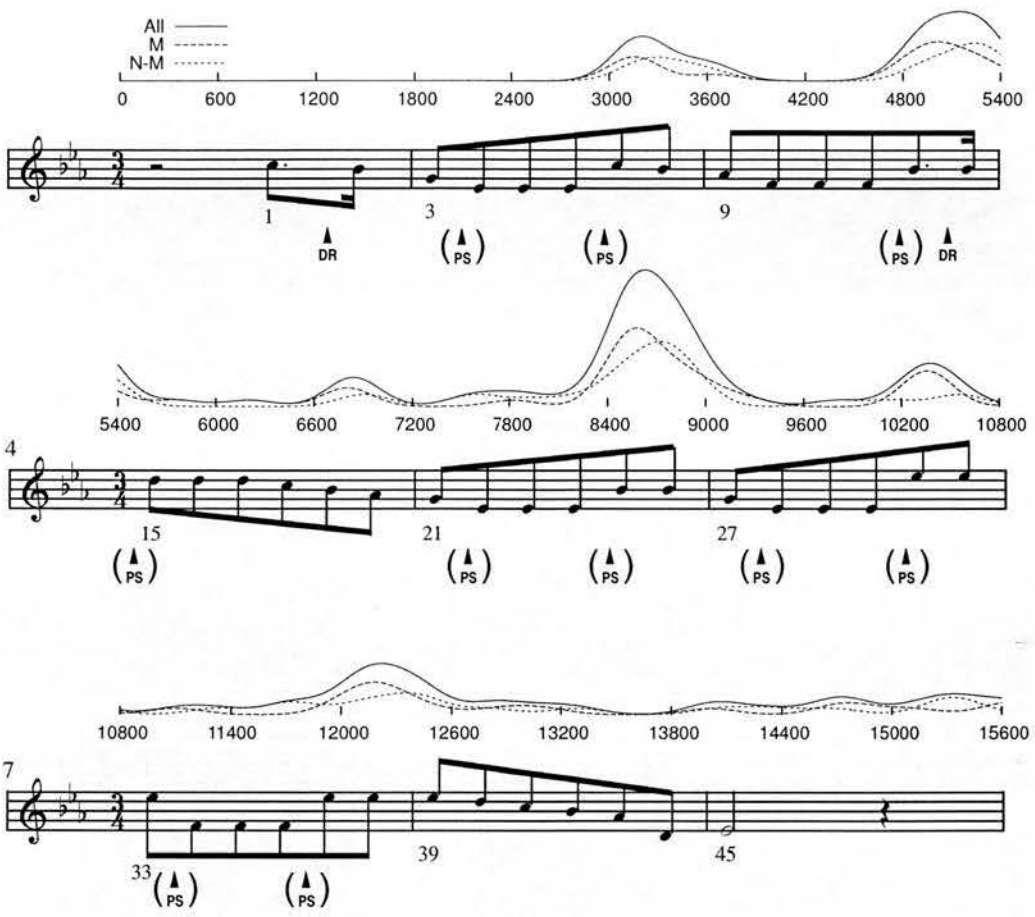


Figure D.6: Comparison between model boundary predictions and listeners' boundaries for melody F0927

t_{LB} (ms)	$pd(t_{LB})(\%)$	$e[-], e[+]$
5620	73	13,14
10430	100	26,27
13220	34	37,38
15720	52	46,47
18270	45	54,55

Table D.6: Listener boundaries selected for melody Q0034. Boundaries depicted with time of occurrence (in ms.), probability density peak value as a percentage of the maximum for the whole melody, and the indexes of the preceding (e[-]) and following event (e[+])

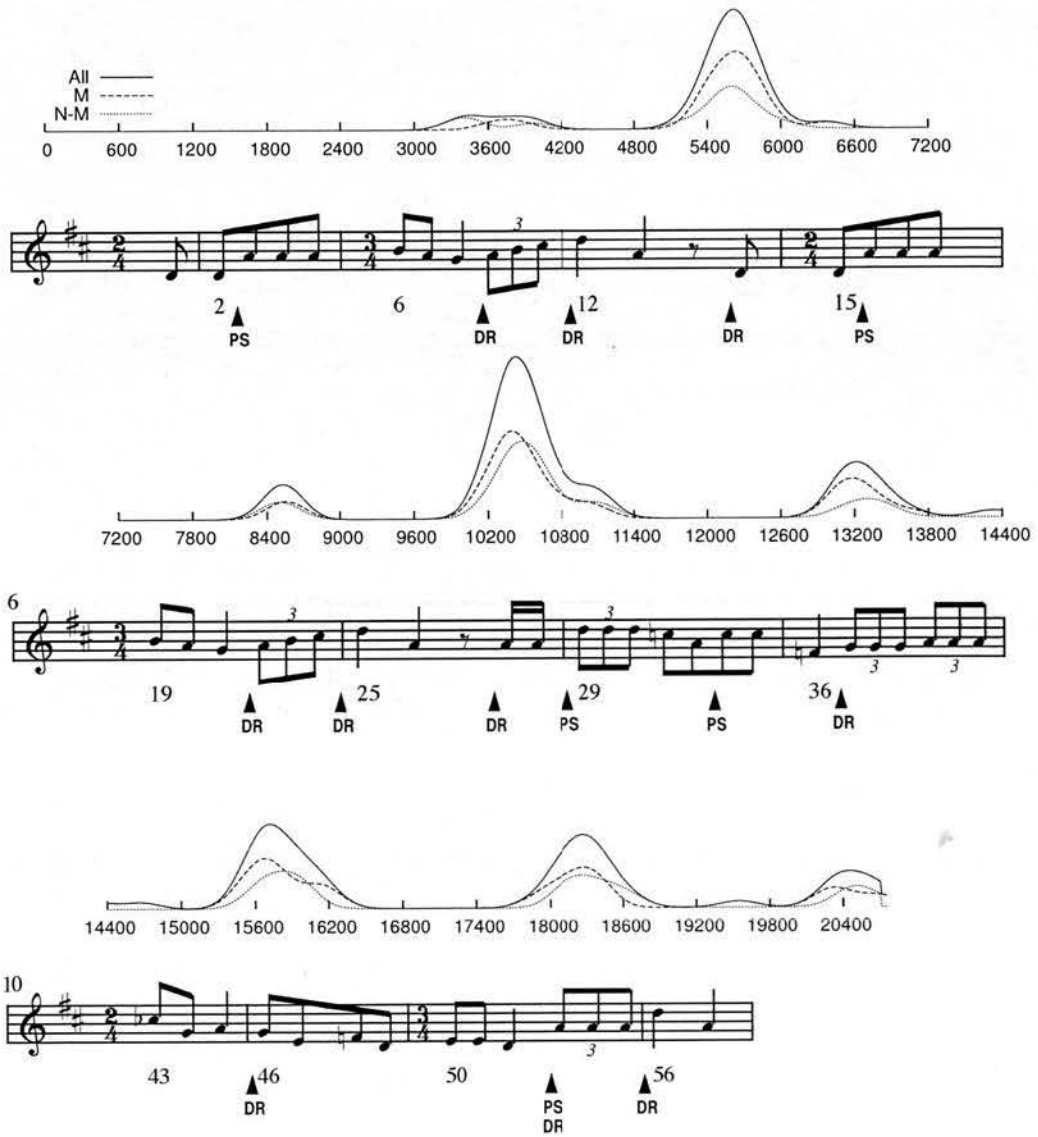


Figure D.7: Comparison between model boundary predictions and listeners' boundaries for melody Q0034

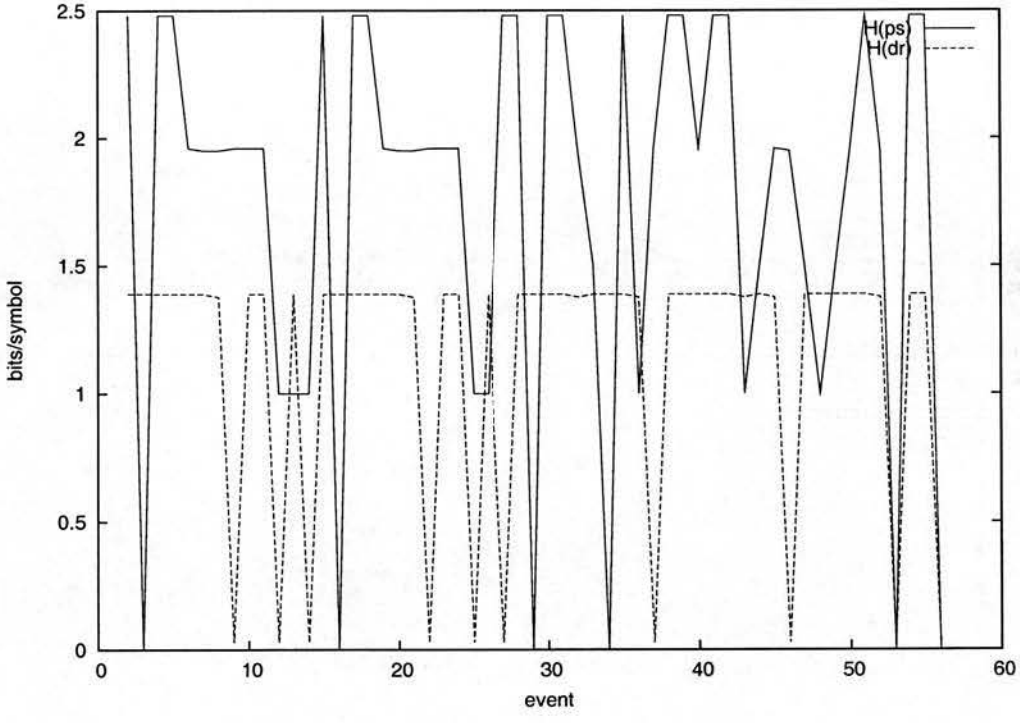


Figure D.8: Folk-song q0034: average entropy $H(c)$ and outcome probability $P(x|c)$

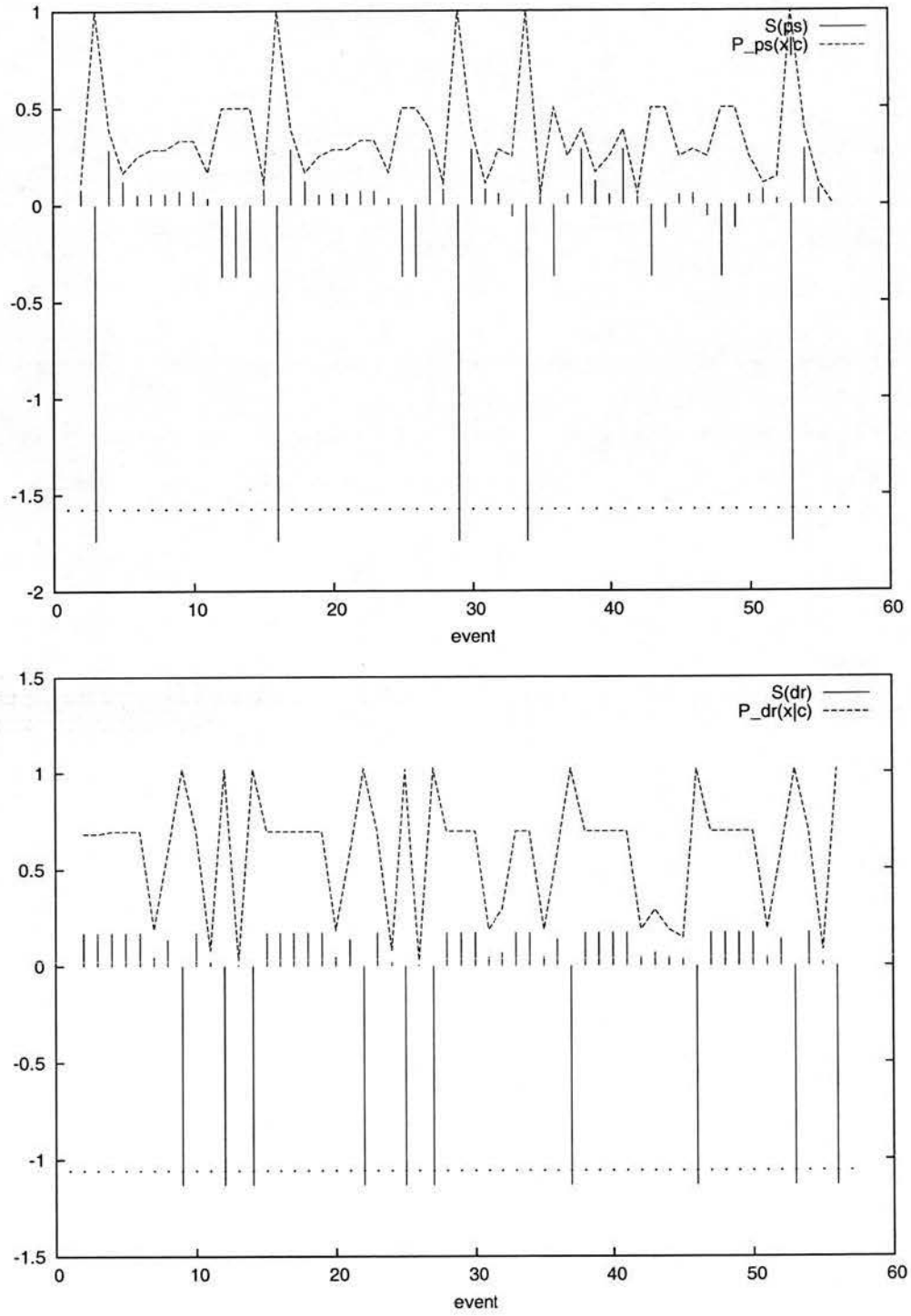


Figure D.9: Folksong Q0034: boundary predictions $S(C)$ and successor probability $P(X|C)$ for features PS and DC. Boundary selection threshold indicated by a dotted line at the bottom of the graph

Appendix E

Supplementary material

This appendix corresponds to the CD which is included with this dissertation. It contains supplementary material including experimental data, sound files, computer code and other related documentation.

CD contents:

Listening study

Puncher: Music Puncher software (java classes & source)

Results: Segmentation data obtained in the listening study

M4

Mixed-Memory Markov Model implementation (java classes & source)

+ documented example on how to use the code.

Melodies

Midis: MIDI files used in the listening study

Events: Event lists of MIDI files

Other

KDE: Kernal Density Estimation software (java classes & source)

Midi Extractor: MAX/MSP MIDI event extractor

Publications

Articles published in the course of this research.

Bibliography

Samer Abdallah and Mark Plumbley. Unsupervised learning for music perception, 1999. In Cambridge Music Processing Colloquium, Cambridge, England. Online at http://www.eee.kcl.ac.uk/-samer/docs/cmc99.ps.zip.

Anna Rita Addessi and Roberto Caterina. Perception of the "macroform" in the *Quartetto per Archi in due Tempi* (1955) by Bruno Maderna. In M. Olivetti Belardinelli, ?, and ?, editors, *Proceedings of the 2nd International Conference "Understanding and Creating Music"*, Caserta, Italy, November 2002. Seconda Università di Napoli, Facoltà di Scienze Matematiche Fisiche Naturali.

Rita Aiello. Can listening to music be experimentally studied? In R. Aiello and J.A. Sloboda, editors, *Musical Perceptions*, chapter 12, pages 273–282. Oxford University Press, 1994.

E. Bigand, F. Lerdhal, and M. Pineau. Deux approches expérimentales des quatre composants de la Théorie Generative de la Musique Tonale. In I. Deliége, editor, *Proceedings of the 3rd International Conference on Music Perception and Cognition*, pages 259–260, Liège, Belgium, 1994. European Society for the Cognitive Sciences of Music.

Emmanuel Bigand. Contributions of music to research on human auditory cognition. In Stephen McAdams and Emmanuel Bigand, editors, *Thinking in Sound: The Cognitive Psychology of Human Audition*, chapter 8, pages 231–277. Oxford University Press, 1993.

W. Birmingham, R. Dannenberg, G. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. Musart: Music retrieval via aural queries. In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR), 2001*, pages 73–81, Bloomington, IN, October 2001.

Rens Bod. Memory–based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 30(1):27–37, 2001.

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Massachusetts Institute of Technology, 1990.

Stewart N. Brown, G.D.A, and N. Chater. Sequence effect in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 28(1):3–11, 2002.

E. Cambouropoulos, T. Crawford, and C.S. Iliopoulos. Pattern processing in melodic sequences: challenges, caveats and prospects. *Computers and the Humanities,* 35(1): 9–21, 2001.

Emilios Cambouropoulos. *Towards a General Computational Theory of Musical Structure.* PhD thesis, University of Edinburgh, 1998.

Emilios Cambouropoulos. The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC'2001),* Havana, Cuba, September 2001a.

Emilios Cambouropoulos. Melodic cue abstraction, similarity, and category formation: A formal model. *Music Perception,* 18(3):347–370, 2001b.

Emilios Cambouropoulos. Musical pattern extraction for melodic structure. In R. Kopiez, A.C. Lehmann, I. Wolther, and C. Wolf, editors, *Proceedings of 5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM5),* pages 134–137, Hannover, Germany, 2003.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modelling. In *Proceedings of the 34th Annual Meeting of the ACL,* 1996.

D. Conklin and I.H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research,* 24:51–73, 1995.

David Cope. *Computers and Musical Style.* Oxford University Press, New York, 1991.

Ian Cross. Review of the the analysis and cognition of melodic complexity: the implication-realisation model, by E. Narmour, univ. of chicago press, chicago, 1992. *Music Perception,* 12(4):486–509, 1995.

L.L Cuddy and C.A. Lunney. Expectancies generated by melodic intervals: Perceptual judgements of melodic continuity. *Perception & Psychophysics,* 57:451–462, 1995.

E. J. Davelaar, Y. Goshen-Gottstein, A. Ashkenazi, H. J. Haarmann, and M. Usher. The demise of short-term memory revisited: empirical and computational investigation of recency effects. *Psychological Review*, 112(1):3–42, 2005.

Irène Deliège. Grouping conditions in listening to music: an approach to Lerdahl and Jackendoff's grouping preference rules. *Music Perception*, 4:325–360, 1987.

Irène Deliège. Similarity in processes of categorization. In *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, SimCat 97*, pages 59–65, University of Edinburgh, Edinburgh, Scotland, 1997.

Irène Deliège. Wagner "alte weise": Une approche perceptive. *Musica Scientiæ*, Special Issue:63–90, 1998.

Irène Deliège. Prototype effects in music listening: An empirical approach to the notion of imprint. *Music Perception*, 18(3):371–407, 2001.

Irène Deliège and Marc Melén. Cue abstraction in the representation of musical form. In Irène Deliège and John Sloboda, editors, *Perception and Cognition of Music*, chapter 16, pages 387–412. Psychology Press, 1997.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (B)*, 39(1):1–38, 1977.

P. Desain and H. Honing. Computational models of beat induction: The rule-based approach. *Journal of New Music Research*, 28(1):29–42, 1999.

P. Desain, H. Honing, H. van Thienen, and W.L. Windsor. Computational modeling of music cognition: Problem or solution? *Music Perception*, 16(1):151–166, 1998.

D. Deutsch. Octave generalization and tune recognition. *Perception & Psychophysics*, 11: 411–412, 1972.

W. Jay Dowling. Melodic contour in hearing and remembering melodies. In R. Aiello and J.A. Sloboda, editors, *Musical Perceptions*, chapter 7, pages 173–190. Oxford University Press, 1994.

W.J. Dowling. Recognition of melodic transformations: Inversion, retrograde and retrograde inversion. *Perception & Psychophysics*, 12(5):417–421, 1972.

W.J. Dowling and A.W. Hollombe. The perception of melodies distorted by splitting into several octaves: Effects of increasing proximity and melodic contour. *Perception & Psychophysics*, 21:60–64, 1977.

T. Eerola, P. Toiviainen, and C.L. Krumhansl. Real-time prediction of melodies: Continuous predictability judgments and dynamic models. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick, editors, *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC 7)*, pages 473–476, Sydney, July 2002.

Réné von Egmond, Dirk-Jan Povel, and Eric Maris. The influence of height and key on the perceptual similarity of transposed melodies. *Perception & Psychophysics*, 58(6): 1252–1259, 1996.

M. Eysenck and M.T. Keane. *Cognitive Psychology: a Student's Handbook*. Psychology Press, 3rd edition, 1995.

Miguel Ferrand, Peter Nelson, and Geraint Wiggins. A probabilistic model for melody segmentation. In *Electronic Proceedings of the 2nd International Conference on Music and Artificial Intelligence (ICMAI'2002)*, University of Edinburgh, Scotland, September 2002. URL `http://www.music.ed.ac.uk/student/pages/pg/mferrand/publications/icmai0%2.ps.Z`.

Miguel Ferrand, Peter Nelson, and Geraint Wiggins. Memory and melodic density: A model for melody segmentation. In Nicola Bernardini, Francesco Giomi, and Nicola Giosmin, editors, *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, pages 95–98, Firenze, Italy, 2003a.

Miguel Ferrand, Peter Nelson, and Geraint Wiggins. Unsupervised learning of melodic segmentation: A memory-based approach. In R. Kopiez, A.C. Lehmann, I. Wolther, and C. Wolf, editors, *Proceedings of 5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM5)*, pages 141–144, Hannover, Germany, 2003b.

R.L. Goldstone. Similarity. In R.A. Wilson and F.C. Keil, editors, *MIT encyclopedia of the cognitive sciences*, pages 763–765. MIT Press, Cambridge, MA, 1999.

R.L. Goldstone and J.Y. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning*, chapter 2, pages 273–282. Cambridge University Press, Cambridge, 2005.

U. Hahn, N. Chater, and L.B.C. Richardson. Similarity as transformation. *Cognition*, 87: 1–32, 2003.

Stephen Handel. The effect of tempo and tone duration on rhythmic discrimination. *Perception & Psychophysics*, 54(3):370–382, 1993.

E. Hannon, Joel Snyder, Tuomas Eerola, and Carol Krumhansl. The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, 30(5):956–974, 2004.

E. E. Hannon and S. P. Johnson. Infants use meter to categorize rhythms and melodies: Implications for musical structure learning. *Cognitive Psychology*, 50:354–377, 2005.

Walter B. Hewlett and Eleanor Selfridge-Field, editors. *Melodic Similarity, Concepts, Procedures and Applications*, volume 11 of *Computing in Musicology*. MIT Press, 1998.

Ludger Hofmann-Engl and Richard Parncutt. Computational modeling of melodic similarity judgments: two experiments on isochronous melodic fragments. URL `http://freespace.virgin.net/ludger.hofmann-engl/similarity.html`. 1998.

S.H. Hulse and S.C. Page. Toward a comparative psychology of music perception. *Music Perception*, 5(4):427–452, 1988.

David Huron. What is a musical feature? Forte's analysis of Brahms's opus 51, no. 1, revisited. *Music Theory On-line*, 7(4), 2001.

F. Jelinek and R. Mercer. Interpolation estimation of Markov source parameters for sparse data. *Pattern Recognition in Practice*, pages 381–397, 1980.

W. Jestead, D. Luce, and D.M. Green. Sequential effect in judgements of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1):92–104, 1977.

C. L. Krumhansl. Effects of perceptual organisation and musical form on melodic expectancies. In Mark Leman, editor, *Music, Gestalt and Computing: Studies in Cognitive and Systematic Musicology*, volume LNCS/LNAI 1317, pages 294–319. Springer-Verlag, Berlin, Germany, 1997.

Carol L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Number 17 in Oxford Psychology Series. Oxford University Press, Department of Psychology, Cornel University, 1990.

Carol L. Krumhansl. Memory for musical surface. *Memory and Cognition*, 19(4):401–411, 1991.

Alexandra Lamont and Nicola Dibben. Perceived similarity of musical motifs: An exploratory study. In *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, SimCat 97*, pages 143–149, University of Edinburgh, Edinburgh, Scotland, 1997.

Alexandra Lamont and Nicola Dibben. Motivic structure and the perception of similarity. *Music Perception*, 18(3):245–274, 2001.

Marc Leman and Albrecht Schneider. Origin and nature of cognitive and systematic musicology: An introduction. In Mark Leman, editor, *Music, Gestalt and Computing: Studies in Cognitive and Systematic Musicology*, volume LNCS/LNAI 1317. Springer-Verlag, Berlin, Germany, 1997.

Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge (Mass.), 1983.

Daniel J. Levitin. Memory for musical attributes. In P.R. Cook, editor, *Music Cognition and Computerized Sound: An Introduction to Psychoacoustics*, chapter 17, pages 209–227. M.I.T. Press, 1999.

H.C. Longuet-Higgins and C.S Lee. The perception of musical rhythms. *Perception*, 11: 115–128, 1982.

Christopher D. Manning and Hinrich Schüttze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass., 1999.

Dominic W. Massaro, Howard J. Kallman, and Janet L. Kelly. The role of tone height, melodic contour and tone chroma in melody recognition. *Journal of Experimental Psychology*, 6(1):77–90, 1980.

D. Meredith, K. Lemström, and G. A. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

Leonard B. Meyer. *Emotion and Meaning in Music*. University of Chicago Press, 1956.

Leonard B. Meyer. *Music, The Arts, And Ideas – Patterns and Predictions in Twentieth–Century Culture*. University of Chicago Press, Chicago, 1967.

Leonard B. Meyer. *Explaining Music: Essays and Explorations*. University of California Press, Berkeley, 1973.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

Eugène Narmour. *The Analysis and Cognition of Melodic Complexity: The Implication-realisation Model*. University of Chicago Press, Chicago, 1990.

Eugène Narmour. *The Analysis and Cognition of Basic Melodic Structures: The Implication-realisation Model*. University of Chicago Press, Chicago, 1992.

Jean-Jacques Nattiez. *Fondements d'une Sémiologie de la Musique*. Union Générale d'Editions, Paris, 1975.

Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38, 1994.

Caroline Palmer and Carol L. Krumhansl. Independent temporal pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1):116–126, 1987.

M. T. Pearce and G.A. Wiggins. Rethinking gestalt influences on melodic expectancy. In S.D. Lipscomb, R. Ashley, R.O. Gjerdingen, and P. Webster, editors, *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC 8)*, pages 439–443, Evanston, IL, August 2004.

Judy Platinga and Laurel Trainor. Memory for melody: infants use a relative pitch code. *Cognition*, 98:1–11, 2005.

L. Pollard-Gott. The emergence of thematic concepts in repeated listening to music. *Cognitive Psychology*, 1:66–94, 1983.

Dan Ponsford, Geraint Wiggins, and Chris Mellish. Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177, 1999.

D. J. Povel and P. Essens. Perception of temporal patterns. *Music Perception*, 2(4):411–440, 1985.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE 77*, volume 77, pages 257–286, 1989.

Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3): 205–229, 1989.

Y. Ben Reis. Simulating music learning: On–line, perceptually guided pattern induction of context models for multiple-horizon prediction of melodies. In *Proceedings of AISB'99 - Symposium on Musical Creativity*, pages 58–63, 1999.

Mark Reybrouck. Gestalt concepts and music: Limitations and possibilities. In Mark Leman, editor, *Music, Gestalt and Computing: Studies in Cognitive and Systematic Musicology*, volume LNCS/LNAI 1317, pages 57–69. Springer-Verlag, Berlin, Germany, 1997.

F. Rieke, D. Warland, R. de R. Van Steveninck, and W. Bialek. *Spikes: exploring the neural code*. The MIT Press, Cambridge, MA, 1997.

Irvin Rock and Stephen Palmer. The legacy of gestalt psychology. *Scientific American*, pages 48–61, December 1990.

Pierre-Ives Rolland and Jean-Gabriel Ganascia. Musical pattern extraction and similarity assessment. In Edurado Reck Miranda, editor, *Readings in Artificial Intelligence*, chapter 7, pages 115–144. Harwood Academic Publishers, 1999.

Frank A. Russo and Lola L. Cuddy. Predictive value of Narmour's principles for cohesiveness, pleasingness, and memory of Webern melodies. In *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC'96)*, pages 439–443, Montreal, 1996.

Nicolas Ruwet. *Langage, Musique et Poésie*. Editions du Seuil, Paris, 1972.

Laurence Saul and Fernando Pereira. Aggregate and mixed-order markov models for statistical language processing. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 81–89, Somerset, New Jersey, 1997. Association for Computational Linguistics.

Lawrence Saul and Michael Jordan. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 37(1):75–87, 1999.

Rebecca Schaeffer, Jaap Murre, and Rens Bod. Limits to universality in segmentation of simple melodies. In S.D. Lipscomb, R. Ashley, R.O. Gjerdingen, and P. Webster, editors, *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC 8)*, pages 439–443, Evanston, IL, August 2004.

H. Schaffrath. *The ESAC electronic songbooks*, volume 9 of *Computing in Musicology*. 1994.

Arnold Schoenberg. Style and idea: selected writings by arnold schoenberg. Faber, London, 1975.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, July and October 1948.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

Bob Snyder. *Music and Memory: An Introduction*. MIT Press, Cambridge (Mass.), 2000.

J. Spencer-Smith and R. L. Goldstone. The dynamics of similarity. *Bulletin of the Japanese Cognitive Science Society*, 4:38–56, 1997.

Christian Spevak, Belinda Thom, and Karin Höthker. Evaluating melodic segmentation. In Christina Anagnostopoulou, Miguel Ferrand, and Alan Smaill, editors, *Music and Artificial Intelligence, Second International Conference, ICMAI 2002, Edinburgh, Scotland, UK, September 12-14, 2002, Proceedings*, volume 2445 of *Lecture Notes in Computer Science*, pages 168–182. Springer, 2002.

David Temperley. *The Cognition of Basic Melodic Structures*. MIT Press, Cambridge, Massachusetts, 2001.

Belinda Thom, Christian Spevak, and Karin Höthker. Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the 2002 International Computer Music Conference (ICMC'02*, Göteborg, Sweden, 2002.

Petri Toivianen and Joel S. Snyder. Tapping to Bach: Resonance-based modeling of pulse. *Music Perception*, 21(1):43–80, 2003.

Laurel J. Trainor and Sandra E. Trehub. A comparison of infants' and adults' sensitivity to western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):394–402, 1992.

A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

Réné van Egmond and Dirk-Jan Povel. Perceived similarity of exact and inexact transpositions. *Acta Psychologica*, 92:283–295, 1996.

C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, Boston, 1979.

Paul von Hippel. Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception*, 17(3):315–327, 2000.

L.M. Ward and G.R. Lockhead. Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, 84(1):23–34, 1970.

M. Wertheimer. Laws of organization in perceptual forms. In Willis D. Ellis, editor, *A Sourcebook of Gestalt Psychology*, pages 71–88. K. Paul, Trench, Trubner & Co., London, 1938. (partial translation published in 1938 of "Untersuchungen zur Lehre von der Gestalt" , II Psychologische Forschung 4, 1923, pp 301–305).

Benjamin W. White. Recognition of distorted melodies. *American Journal of Psychology*, 73:100–107, 1960.

Geraint Wiggins. A review on (Hewlett and Selfridge-Field, 1998). *Musicae Scientiæ*, 3 (2):279–283, 1999.

I. H. Witten, L. C. Manzara, and D. Conklin. Comparing human and computational models of music prediction. *Computer Music Journal*, 18(1):70–80, 1994.

MEMORY AND MELODIC DENSITY: A MODEL FOR MELODY SEGMENTATION

Miguel Ferrand, Peter Nelson *Geraint Wiggins*

University of Edinburgh
Scotland, UK
{mferrand,pwn}@music.ed.ac.uk

City University
London, UK
geraint@soi.city.ac.uk

ABSTRACT

We present a memory-based model for melodic segmentation based on the notion of melodic density. The model emphasises the role of short-term memory and time in music listening, by modelling the effects of recency in the perception of boundaries. We describe the model in detail and compare it with Cambouropoulos' Local Boundary Detection Model for a series of melody examples. First results indicate that this new model is more conservative, as it generates fewer total boundaries but preserves most boundaries that coincide with the limits of recurring patterns.

1. INTRODUCTION

It is known that listeners identify segmentation boundaries when abstracting musical contents. The ability to partition a melody in several segments provides a structural description of the piece of music. Thus, segmentation can be seen as a pre-processing stage for other tasks such as pattern discovery or music search.

Pattern finding algorithms, in particular, are known to be computationally expensive, and therefore can benefit from a reduction of the initial search space. A low-level segmentation can provide an efficiency gain by pre-processing a melodic sequence, and generating an initial set of boundaries which may be used as markers for pattern search [1]. One such method is The Local Boundary Detection Model (LBDM) [2], a segmentation model that identifies discontinuities in a melodic surface based on Gestalt principles of perception. The LBDM is an essential reference amongst segmentation algorithms, mostly due to its simplicity and generality [3, 2]. As the author emphasises, the LBDM is not a complete model of grouping in itself, as it relies on complementary models (i.e. pattern similarity) to select the most relevant boundaries. Although in that context this may not be considered a weakness of the model, excessive boundary generation may become a disadvantage if we intend to use the LBDM in isolation, and when segmentation is to be used as a reliable data reduction technique.

The LBDM has a fairly short memory as it considers at most 4 consecutive events at a time. As a consequence, there is limited interaction between neighboring boundaries and sometimes small "oscillations" can be identified as salient boundaries. This type of limitation has also been referred to by Lerdahl & Jackendoff in their Generative Theory of Tonal Music [4].

Research on auditory perception and memory has underlined the influence of time in the perception of differences and in the establishment of temporal relations in sequential processes. Studies have shown that listeners retain auditory information for some time, even after the end of stimulation [5]. This means that several past (although relatively recent) stimuli may draw the listener's attention, and may be retained as the actual most recent and prominent stimuli. Some researches have suggested that listeners perceive a musical surface by focusing on successive zones, that can be viewed as a "sliding window" along the musical piece [6]. The size of this window (determined by short-term memory restrictions) should limit the amount of musical material that can be looked back on when processing a melodic sequence. Within this time window, recency effects are likely to apply, as documented in [7, 8].

2. THE LBDM

The LBDM calculates a boundary profile for a melody, using Gestalt-based identity–change and proximity–difference rules, applied to several parameters describing a melody. The refined version of this algorithm [2] takes as input a melodic sequence converted into several independent parametric interval profiles $P_k = [x_1, x_2, ...x_n]$ where $k \in \{pitch, ioi, rest\}, x_i \geq 0$ and $i \in \{1, 2, ...n\}$. A $Change$ rule assigns boundaries to intervals with strength proportional to the degree of change between neighboring consecutive interval pairs. Then a $Proximity$ rule scales the previous boundaries proportionally to the size of the intervals.

The strength of the boundaries at each interval x_i is given by the following,

$$s_i = x_i \times (r_{i-1,i} + r_{i,i+1}) \tag{1}$$

where

$$r_{i,i+1} = \begin{cases} \frac{|x_i - x_{i+1}|}{x_i + x_{i+1}} & x_i + x_{i+1} \neq 0 \wedge x_i, x_{i+1} \geq 0 \\ 0 & x_i = x_{i+1} = 0 \end{cases}$$

For each parameter k a sequence s_k is calculated, then all sequences are normalised and combined in a weighted sum to give the overall boundary strength profile. The suggested weights for the 3 different parameters are $w_{pitch} = w_{rest} = 0.25$ and $w_{ioi} = 0.5$ (see [9] for an overview on the behavior of the LBDM with different parameter tunings). The local peaks in the resulting boundary profile indicate local boundaries in the melodic sequence. A threshold must be defined a priori, above which, a peak is identified as a boundary. For additional details on the implementation of the LBDM the reader is referred to [2].

3. MELODIC DENSITY SEGMENTATION MODEL

We now describe a new model for melodic segmentation which identifies segmentation boundaries as perceived changes in melodic

Table 1: $Order$ and $recency$ of pitch intervals for a sequence of events. Intervals are in semitones.

e_{i-3}	e_{i-2}	e_{i-1}	e_i	$event$
53	50	50	48	$pitch$
				$order(n)$
	3	0	2	1
		3	2	2
			5	3
...	2	1	0	$recency(m)$

density. We will designate this model as Melodic Density Segmentation Model (MDSM). In contrast with the LBDM, that measures the accumulated boundary strength and identifies local maxima, the MDSM calculates the accumulated melodic cohesion between pitch intervals, and then identifies local minima (i.e. points of low melodic density) as local boundaries. This new segmentation method also incorporates a short-term memory window and models the effects of recency with an attenuation function.

Before a formal description of the model is presented, some of its characteristics and underlying assumptions must be explained.

It is conjectured that pitch intervals may be formed (and perceived) between all notes occurring over an interval of time (short term memory window) and not just between consecutive notes. In Table 1 a short sequence of 4 midi notes is depicted together with the pitch distances between all pairs of events. The order of an interval determines the distance between the present and previous event considered. Thus, an interval of order k with respect to a given event e_i is denoted by (e_{i-k}, e_i). For example, from table 1 intervals (e_{i-1}, e_i) and (e_{i-2}, e_{i-1}) have order 1, intervals (e_{i-2}, e_i) and (e_{i-3}, e_{i-1}) have order 2, etc...

Recency effects apply in two different ways. The higher the order of an interval, the greater the temporal separation between the events, and therefore the weaker the perceived link between the two. On the other hand, more recently formed intervals have a stronger contribution to the melodic cohesion of the sequence than earlier formed ones. The recency of an interval with respect to an event e_i is given by the time that separates e_i and the latest event of the two that constitute the interval. These two factors are combined to determine the overall contribution of each interval at any given moment in time. In Table 1, recency is indicated in the bottom row. Increasing values of recency express less recent intervals. Let's consider here for simplicity, that all events in the previous example have equidistant on-set times and equal duration. Then intervals (e_{i-2}, e_i) and (e_{i-2}, e_{i-1}) will have equivalent contribution, since the former is an interval of order 2 (meaning that events are separated by 2 duration units) but with recency 0, and the latter has order 1 but recency 1 (meaning that the interval is separated from the reference event e_i by 1 duration unit).

The melodic cohesion of an interval is defined here to be proportional to the frequency of occurrence of that interval in the interval framework associated with the melody being analysed. Later, we will discuss in more detail how these interval frequencies are obtained.

A short–term memory window determines the span of recent events that can form intervals. The size (duration) of this window is fixed. The tempo of the piece will determine the number of recent events that can be recalled and influence the perception of a boundary.

We can now formalise the notion of melodic density (MD) as the weighted sum of the contributions of all intervals occurring over a period of time determined by the memory window. So given a sequence of N events $(e_1, ..., e_N)$ representing a melodic sequence the melodic density d_i at event i, is defined as:

$$d_i = \sum_{m=0}^{t_i - t_{i-m} < M} \sum_{n=1}^{t_i - t_{i-m-n} < M} f(r_i(m,n)) \cdot a_i(m,n) \quad (2)$$

where $f(r)$ is a function that returns the frequency of an interval, and $f(r) \in [0,1], r_i \in 0, 1, ...12$, and $r_i(m,n) = |p_{i-m} - p_{i-m-n}|$ denotes a pitch interval in semitones, where p_k denotes the MIDI pitch of event e_k, and

$$a_i(m,n) = (1 - \frac{t_i - t_{i-m-n}}{M})^2 \quad (3)$$

is the attenuation function, where t_i denotes the onset time of event e_i, and M is the duration of the memory window (in seconds). It is worth noting that a Gestalt-based principle of proximity is encapsulated in the attenuation function, as this will return values closer to 1 for recent and low-order intervals, and values closer to 0 for remote and high-order intervals.

Finally, boundaries are indicated by local minima in the melodic density profile obtained from Equation 2.

4. EXPERIMENTS AND RESULTS

To assess the behavior of the model we used both the LBDM and the MDSM on a set of melody examples. For each of the examples we also obtained a pattern boundary profile, which indicates the location of recurrent patterns within the melodic sequence (see [1] for details).

The interval frequencies given by function f were obtained from the combined frequencies of intervals that occurr in major and minor scales. This major-minor framework is described by Camboroupoulos in his General Pitch Interval Representation (GPIR) [1]. The memory window M was set to 4 seconds.

Table 2 summarises the boundary counts for each melody, including pattern boundaries and the segment boundaries generated by both the LDBM and the MDSM. A boundary is marked correct if its location coincides with a pattern boundary, with a tolerance of +/-1 event. A threshold of 70% was adopted to filter only the most prominent peaks from the boundary profiles.

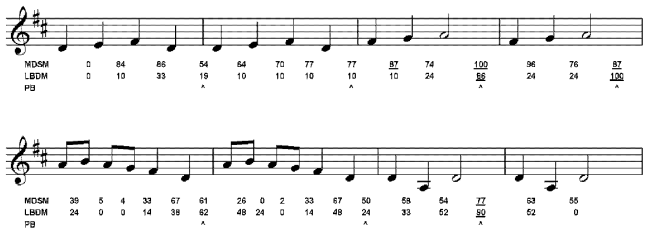


Figure 1: Normalised MDSM and LBDM boundary profiles for melody number 2 (Frere Jacques). Underlined values indicate selected peaks. Pattern Boundaries(PB) are indicated in the bottom row

Table 2: Results obtained for 7 melodies, showing the total no. of pattern boundaries (PB), and for both the LBDM and MDSM: total no. of pattern boundaries found ($_{fnd}$), no. of pattern boundaries not found ($_{notfnd}$) and no. expurious boundaries found ($_{ex}$)

		LBDM			MDSM		
Melody	PB	$_{fnd}$	$_{notfnd}$	$_{ex}$	$_{fnd}$	$_{notfnd}$	$_{ex}$
1. L. Row	5	5	0	0	5	0	0
2. Frere J.	7	3	4	0	5	2	0
3. Twinkle	5	5	0	2	4	1	1
4. Y.Doodle	5	5	2	3	5	0	2
5. L'H.Arme	9	8	1	0	9	0	0
6. Mozt.Gm	6	6	0	14	6	0	3
7. Beet.9th	9	9	0	0	9	0	0
Total	46	39	7	19	43	3	6

Table 3: F-measure for the LBDM and MDSM

Model	P	R	F
LBDM	0.85	0.67	0.75
MDSM	0.93	0.88	0.91

In total the LBDM generated 58 boundaries against only 49 by the MDSM. From the analysis of Table 2 it may be observed that both models find approximately the same number of pattern boundaries, but the MDSM is more conservative, generating only 6 excessive boundaries, against the 19 of the LBDM. In the melodies where excessive boundaries where found, the MDSM always register a lower count. However It must be noted that melody number 6 alone (theme of Mozart's Symphony in Gm) is responsible for the majority of the excessive boundaries generated by the LBDM. For a numerical comparison between the performance of both models the F-measure [10] was used. The F-measure is given by the weighted harmonic mean of $Precision(P)$ and $Recall(R)$.

$$F_{measure} = 2 \times \frac{P \times R}{P + R} \tag{4}$$

where

$$P = \frac{PB_{fnd}}{PB_{fnd} + PB_{notfnd}}, R = \frac{PB_{fnd}}{PB_{fnd} + PB_{excessfnd}} \tag{5}$$

In table 3 we can see that although the MDSM only has a slightly higher $Precision$, it has a significantly higher $Recall$ resulting in a higher value of F.

In Figure 1 we show the boundary profiles of both models together with the score of melody no. 2 (Frere Jacques). For ease of comparison, the melodic density profile of the MDSM has been inverted [1] and normalised in the range 0-100%. From this example it seems clear that some of the boundaries generated by the LBDM were eliminated due to the 70% selection threshold, although smaller peaks can be found in the vicinity of the pattern boundaries that were missed.. An adjustment of the selection threshold to considerably lower values, will result in a significant increase of the number of peaks that are extracted, and consequently in an increase of the number of spurious boundaries. On the other hand, we would expect that an increase of the selection threshold would increase the selectivity of the model. In Figure 2 we can observe that this is not always the case. Most of the peaks of the LBDM profile have values over 80% or even 90%, thus making the elimination of the excessive boundaries difficult to achieve only by adjusting the selection threshold. The example of Figure 2 highlights also that most of the boundaries "filtered" by the MDSM are not coincident with pattern boundaries.

[1] recall that for the MDSM boundaries are obtained from the lower peaks on the profiles

5. DISCUSSION

The boundary selectivity reported on the MDSM, results partially from the propagation of the intervals over a time window creating a "smoothing" effect. However this effect can be also a drawback of this approach. In some cases, boundaries can be shifted forward or prolonged due to a slower decay of the melodic density function. This is visible in Figure 1 where the boundary peak after the third measure is followed by a significantly slow decay of the MDSM values (specially when compared with the sharp drop on the LDBM profile), until it meets the following peak. This may have an impact on the accuracy of the boundary locations, in particular when matched without tolerance, against pattern boundaries.

Although tempo was kept constant in this study, the MDSM is robust to small changes in tempo. This is mainly due to the discrete nature of the events, combined with a memory window of fixed size. For example, with a tempo of crotchet=60, a memory window of 5 seconds would include 5 crotchets (or the equivalent in duration), and an increase of the tempo to crotchet=72 would be necessary to include an additional crotchet in the calculations. Few studies have addressed the effects of changes in tempo in music perception [11]. Although the present model was designed to account for changes in tempo, a systematic evaluation of these effects has not yet been included. For such analysis we may require that listeners be tested on the effects of changes in tempo to provide data to be compared with the model.

The choice of the attenuation function (a decaying polynomial), is the result of preliminary experiments with the algorithm, where several decaying functions were examined. However, it must be said, the differences were not conclusive. It seems intuitive that, in general, less recent notes have a smaller contribution to the melodic cohesion of a sequence, than more recent ones. However, to the best of our knowledge, there is no theoretical or experimental evidence to support the choice of a specific memory decaying function.

As mentioned previously, interval frequencies were obtained from the combined statistics of interval counts from major and minor scales. Since one of the motivations of this work is to devise a model that can segment melodies without any domain specific knowledge, we propose that these frequencies may be acquired from a music corpus that is representative of the melodies being analysed. This idea is supported by several studies, some of which were carried out outside the western musical culture, that report, for example, the prevalence of small melodic intervals in melodic lines [7, 12]. If indeed the melodic preferences of a particular musical culture are reflected in the musical material, it seems rea-

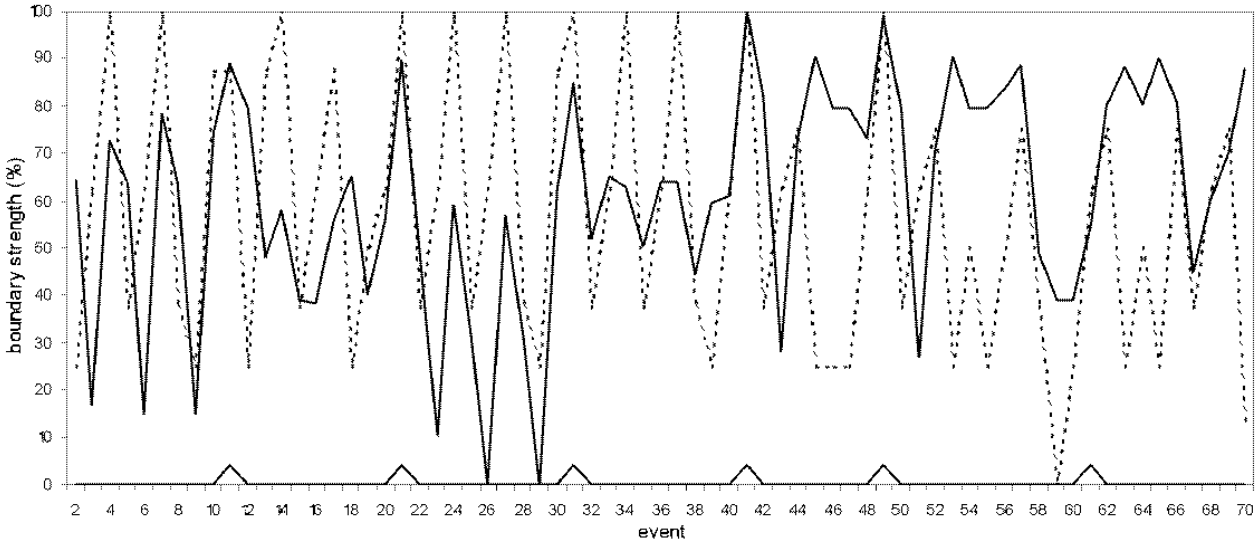


Figure 2: Boundary profiles obtained with LBDM (dotted line) and MDSM (solid line) for melody no. 6 (theme of Mozart's Symphony in Gm). Pattern boundaries are indicated by arrows at the bottom of the chart.

sonable to reverse this process, by using implicit intervalic information to interpret the musical material.

6. CONCLUSIONS

We presented the MDSM, a memory-based melodic segmentation algorithm based on the concept of melodic density. We compared this algorithm with the LBDM, for a set of melody examples. It was shown that in general the MDSM has higher selectivity than the LBDM, generating fewer total boundaries but preserving most boundaries indicated as pattern boundaries. This suggests that the MDSM may be used successfully as a pre-processing method for pattern finding algorithms, providing additional reduction of the search space without the cost of eliminating many candidate pattern boundaries.

The contribution of this new approach lies in the way it incorporates pitch and time, and in particular in the use of tempo as a parameter together with a short-term memory window, thus seeking a more cognitively realistic approach to melodic segmentation.

7. ACKNOWLEDGMENTS

This research was funded by EPSRC research project GR/N08049/01

8. REFERENCES

[1] Emilios Cambouropoulos, *Towards a General Computational Theory of Musical Structure*, Ph.D. thesis, University of Edinburgh, 1998.

[2] Emilios Cambouropoulos, "The local boundary detection model (lbdm) and its application in the study of expressive timing," in *Proceedings of the International Computer Music Conference (ICMC'2001)*, Havana, Cuba, September 2001.

[3] Emilios Cambouropoulos, "Melodic cue abstraction, similarity, and category formation: A formal model," *Music Perception*, vol. 18, no. 3, pp. 347–370, 2001.

[4] Fred Lerdahl and Ray Jackendoff, *A Generative Theory of Tonal Music*, M.I.T. Press, Cambridge (Mass.), 1983.

[5] M. Eysenck and M.T. Keane, *Cognitive Psychology: a Student's Handbook*, Psychology Press, 3rd edition, 1995.

[6] Emmanuel Bigand, "Contributions of music to research on human auditory cognition," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, Stephen McAdams and Emmanuel Bigand, Eds., chapter 8, pp. 231–277. Oxford University Press, 1993.

[7] Carol L. Krumhansl, *Cognitive Foundations of Musical Pitch*, Number 17 in Oxford Psychology Series. Oxford University Press, Department of Psychology, Cornel University, 1990.

[8] Ludger Hofmann-Engl and Richard Parncutt, "Computational modeling of melodic similarity judgments: two experiments on isochronous melodic fragments," http://freespace.virgin.net/ludger.hofmann-engl/similarity.html, 2000.

[9] Belinda Thom, Christian Spevak, and Karin Höthker, "Melodic segmentation: Evaluating the performance of algorithms and musical experts," in *Proceedings of the 2002 International Computer Music Conference (ICMC'02*, Göteborg, Sweden, 2002.

[10] C. J. Van Rijsbergen, *Information Retrieval, 2nd edition*, Butterworths, London, Boston, 1979.

[11] Stephen Handel, "The effect of tempo and tone duration on rhythmic discrimination," *Perception & Psychophysics*, vol. 54, no. 3, pp. 370–382, 1993.

[12] Paul von Hippel, "Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals," *Music Perception*, vol. 17, no. 3, pp. 315–327, 2000.

UNSUPERVISED LEARNING OF MELODIC SEGMENTATION: A MEMORY-BASED APPROACH

Miguel Ferrand[1], *Peter Nelson*[1], *Geraint Wiggins*[2]

[1]University of Edinburgh, Scotland, UK
{mferrand,pwn}@music.ed.ac.uk

[2]City University, London, UK
geraint@soi.city.ac.uk Roman

ABSTRACT

In this paper we propose a memory-based model for melodic segmentation. We argue that the perception of segment boundaries is related to the unpredictability of certain musical features and that feature salience can be learned from a corpus of non-annotated musical data. We describe the implementation of this model and how it uses the acquired information to predict the location of segment boundaries for a given melody. Finally we present some experimental results to show that the model has a significant predictive power regarding the location of segment boundaries, when compared with segment boundaries obtained with listeners.

1. BACKGROUND

When listening to a piece of music, listeners often identify distinct sections or segments within the piece. Music segmentation is recognised as an important step in the abstraction of musical contents and researchers have attempted to explain how listeners perceive and identify the boundaries of these segments.

Existing theories on music segmentation have employed Gestalt principles to identify discontinuities and create groupings between musical events [1,2]. The perception of parallelism and similarities are also known to influence the listener to relate different passages within a musical piece. In fact many Gestalt-based approaches rely on higher-level grouping rules or similarity functions to identify larger scale segment boundaries.

Often, it is suggested that Gestalt principles operate independently of the listeners musical knowledge. When familiarised with a certain musical repertoire, listeners memorise recurrent features in the music and use this knowledge to carry out musical analytical tasks. Empirical evidence has shown that large sections of a musical piece can be recalled by listeners based on the recurrence of small musical cells [3], which act as markers within the piece. Leonard Meyer in his theory on music expectation also outlines the importance of learning in music understanding and relates expectation with information theoretical notions such as entropy [4].

The entropy (or unpredictability) associated with the occurrence of a musical event can change its prominence and hence make it salient to the listener, within a sequence of events. The notion of salience has also been referred as being associated with features that present intra-textual or inter-textual distinctiveness [5]. Probabilistic methods have been widely used to acquire regularities in large sets of data, with many successful applications in natural language and speech processing [6]. Some of these methods have migrated into the music domain however, probabilistic modelling has been used mostly for music prediction and generation [7-9] and seldom to model musical analytical tasks [10] or listening behaviour.

2. AIMS

We seek the development of a system to learn and perform melodic segmentation in an unsupervised way. Learning from raw musical data (without annotations) and avoiding the use of *a priori* musical knowledge or musical rules are central motivations of the present work.

Applications for automatic music segmentation include the support to other analytical methods such as music search and pattern finding. The segments found can set the initial search points within a large piece, thus providing a reduction of the initial search space for complex algorithms.

3. A MODEL FOR MELODIC SEGMENTATION

We propose the implementation of a memory-based model to automatically predict the location of segmentation boundaries in a melody. The three main aspects of this model are described in this section. The first relates to the input of the model and deals with the representation of melodic information. The second and central part of the model is the feature learning module, which is implemented based on Markov models. The third relates to the output of the model and describes how it generates predictions about the location of segment boundaries, given a test melody.

3.1. Melody Representation

Music can be seen as a temporal process where sound events unfold in time. In this work, melody information is converted directly from a Midi source into an event-based symbolic representation including the pitch, duration and the inter-onset interval between events. From these basic attributes we obtained two additional melodic features:

- Pitch step (PS): the interval distance between consecutive events (in semitones).
- Duration ratio (DR): the ratio between the duration of consecutive events.

In Table 1 we show the melodic representation for an extract of Debussy's Syrinx, together with derived features PS and DR. For practical reasons the original DR values (in parenthesis) were converted into a logarithmic scale.

No.	Pitch	Onset	Dur	PS	DR
1	82	3998	1000	-1	-5 (0.14)
2	81	4998	143	+2	-1 (0.80)
3	83	5141	115	-3	6 (8.70)
4	80	5256	1000	-1	-5 (0.17)
5	79	6256	167	+2	0 (0.99)
6	81	6423	166	-2	1 (1.64)
7	78	6589	273	-3	0 (1.18)
8	77	6862	322	…	…

Table 1 – Melody representation and two derived features.

These two features have the advantage of representing melodic information in a relative manner, thus avoiding the use of absolute pitch values or absolute durations. The latter is particularly important since in an expressive (non-mechanical) performance the durations of Midi events do not correspond exactly to the notated durations.

3.2. Feature Learning with Markov Models

Markov models are typically constructed from statistics obtained from a large corpus of data (usually referred to as the training corpus) using the co-occurrences of adjacent symbols to determine the probabilities of sequences of symbols.

An nth order n-gram model (a class of Markov models) assumes that the probability of occurrence of a symbol depends on the prior occurrence of *n-1* other symbols. Given a sequence $s=w_1...w_l$ of length l , the probability $P(s)$ is given by,

$$P(s) = \prod_{i=1}^{l} P(w_i \mid w_{i-1},..., w_{i-n+1}) \quad (1)$$

If the training corpus is small and the order of the model is high, longer sequences will have relatively lower counts, resulting in less accurate probabilities. Independently of the size of the training corpus, it is unlikely that all possible symbol sequences will occur. This becomes a problem if, when computing probabilities using Equation 1, some of the terms in the product have zero probability.

Another disadvantage of n-gram models is that their size increases rapidly with an increase in their order since we may need to store the probabilities of all combinations of fairly long sequences.

Next, we describe Mixed-order Models, which were used in the present work to overcome the increased order and data sparseness problems.

Mixed-order Markov Models

Mixed-order Markov Models (MMM) provide a representation of higher-order models by combining several lower order models [11]. Thus an nth order model over a random variable S (with k possible values) can be expressed as:

$$P(w_i \mid w_{i-1},..., w_{i-n}) = \sum_{\mu=1}^{n} \phi(\mu) a^{\mu}(w_i \mid w_{i-\mu}) \quad (2)$$

where $a^{\mu}(w_i|w_{i-\mu})$ is a $k \times k$ transition matrix containing the probabilities of the occurrence of a symbol at position i given the occurrence of a symbol at position $i-\mu$,

The mixing coefficients $\phi(\mu)$ are estimated using an iterative procedure, using the initial counts in the transition matrix. Due to space restrictions we omit here the description of this procedure. For a detailed explanation of MMMs and parameter estimation methods the reader is referred to [11,12].

The MMM is trained with all pairwise dependencies found in the feature sequences generated from the training set and then the corresponding mixing coefficients are estimated.

2.3 Entropy and Boundary Prediction

Following our initial assumption, we propose that some segmentation boundaries are likely to occur close to accentuated changes in entropy, associated with some melodic features.

Shannon [13] showed that one of the ways of measuring the quantity of information of a particular message is to determine its unpredictability or entropy. We can determine the entropy associated with a given context c as,

$$H_c = \sum_{\forall w} P(w \mid c) \log_2 P(w \mid c) \quad (3)$$

where w denotes all symbols that can be successors of the context c. Context c is a sequence of size *n-1*, where n is the order of the model. Conditional probabilities are obtained from Equation 2 and will reflect the statistics of the training set.

Entropy vectors are then calculated by taking all the successive context sequences from the feature vectors of the target melody. As mentioned earlier we are interested only in more prominent entropy changes across the melody. For every entropy vector we first determine the mean and standard deviation. Then all values outside the standard deviation are filtered from the vectors. Finally, from the remaining values in the vector, we considered only those that register a contiguous low-high or high-low variation with respect to the mean.

3. RESULTS

The experimental part of this work has two components. The first is an empirical study on melodic segmentation carried out with listeners on some melodies. The aim of this study was to collect segmentation information from a real listening experience and to provide comparison data for our computational model of melody segmentation. In the second part we used our computational model with some of the examples provided to the listeners, to predict the locations of the segment boundaries.

3.1. A Listening Study

A total of 48 subjects took part in this listening study. Participants were all 3rd/4th-year undergraduate or postgraduate

students, split between musically trained and non-musically trained subjects.

The set of melodies used in this study included 3 folk songs from the Essen Folk Song Collection (initiated by Prof. H. Schaffrath), 2 melody excerpts from Mozart Piano Sonatas and Debussy's Syrinx. All melodies were provided as deadpan MIDI files, with the exception of Syrinx, which was obtained from an expressive performance (performed by Peter-Jan van Dijk), thus including ornaments, tempo fluctuations and dynamics.

For each melody subjects had two familiarisation auditions, a trial segmentation audition and a final segmentation audition. Data collection was performed by a computer program designed to guide the listeners through the whole listening session. Listeners were able to indicate a segment boundary by pressing the mouse button, while the melody was being played. To minimise the effects of priming, the program also guaranteed that no two listeners heard the melodies in the same order.

The segment locations (time stamps) collected from the listeners were later synchronised with the MIDI data to associate them with the events in the melody. A boundary is matched with an event if it occurs between the onset times of that event and the next. In Figure 1 we show the histogram of segment boundary counts per event for Syrinx, for all subjects. The analysis of the separate boundary histograms for musician and non-musician subjects indicates that the differences between the two are not significant. This relatively low influence of the factor musical training in a segmentation task has previously been reported in [14].

Observing the graph of Figure 1 it is clear that listeners agreed on several segmentation locations, within the melody. The graph, also suggests that there is a delay or anticipation in some of the subjects' responses, particularly visible around the main segment boundaries at events 80, 112, 139 and 252. There are also areas in the graph that show a considerable number of responses that span over a fairly large number of consecutive events (e.g. 47-50 or 303-306).

3.3. Automatic Segmentation

We now look at the results obtained with our segmentation model on Syrinx, the larger melody of the study set. This melody was used both as the training set and the test set.

In Figure 2 we plot the outstanding entropy transitions (white markings) for PS and DR, overlapped with the boundaries indicated by the listeners (we will refer to the latter as *L-boundaries*). In some cases the entropy variations stretch across more than two consecutive events and these, similarly to the *L-boundaries*, are depicted as several overlapped markings.

We considered that a prediction is correct if it indicates an existing *L-boundary* location within a distance of ±1 event. From a total of 14 *L-boundaries* considered, 11 were predicted correctly by the model (5 from *H(PS)* and 6 from *H(DR)*). The model generated also 5 excessive boundaries, 3 from *H(PS)* and 2 from *H(DR)*. Excessive boundaries are those that have no correspondence with any of the *L-boundaries*.

3.4. Discussion

After analysing the nature of the boundaries predicted by the model we observe that most of them can be explained by Gestalt-based principles of proximity both in the pitch and time domains. This idea is reinforced by the fact that most *L-boundaries* coincide with the location of breath marks (see Figure 2), and these often follow longer notes or large pitch intervals marking the end of phrases. Nevertheless, it is a fact that no rules were previously provided to the model, so they have in fact been derived from the data and reflected on the results.

Although we generated models of up to order 6, the weighting coefficients of Equation 2 show that the MMM approximation for this particular melodic data is equivalent to a model of order between 2 and 3. This means that patterns acquired by the model involve at most 3 to 4 events. This imposes a limit on the discovery of pattern similarities. Nevertheless, boundaries at 15, 252 and 268 were predicted based on the rhythmic re-ocurrence of the opening motif of this piece (represented in Table 1). Boundaries 15 and 252 were two of the most voted by listeners. Boundary 268 also marks the beginning of the same rhythmic motif, but was not selected by the listeners. This suggests that although a low order model cannot store large patterns, smaller partial patterns can be retained as indexes of longer ones. Some theories have argued for the prototypical nature of parallelism and have shown that patterns are often remembered by the repetition of smaller cells, often their initial section [16].

The length of Syrinx seems to provide enough redundant information to highlight most of its recurrent features, but not enough to prevent the model from being fairly sensitive to the less frequent ones.

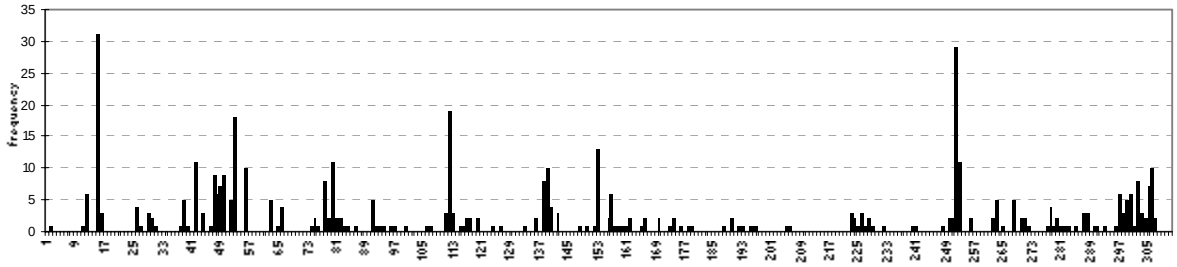


Figure 1 – Histogram of segment boundaries (for all subjects) for Syrinx. Bins correspond to Midi events.

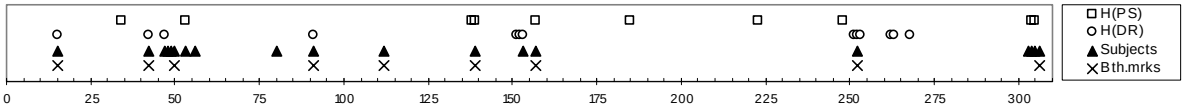


Figure 2 – Segment boundary locations indicated by listeners, notated breath marks and boundaries predicted by the model.

For example, pitch intervals of 1, 2 or 3 semitones are very frequent throughout the whole melody. In comparison most other intervals will seem very improbable, and thus will be responsible for large variations in entropy.

The use of relative measurements for the melodic features used to train the model, increased the redundancy of the data, and to some extent can be seen as a form of representing approximate similarity. It is remarkable that the majority of the *L-boundaries* could be predicted only with the information contained in this one piece. However, for very short melodies this would not be possible due to the lack of redundant information. In the following stage of this research we plan to train the model with a set of melodies and then use a target melody not included in the training set, but that is somehow represented by the training set. More specifically, we plan to use a subset of songs from the Essen Database as our training set. Then use will take the melodies used in the listening study as our test set, and re-evaluate the ability of the model to predict the boundary locations.

5. CONCLUSIONS

We presented a memory-based model of music learning and melodic segmentation. The model relates feature salience with expectation and uses entropy measurements to evaluate the occurrence of pitch and time-based melodic features.

We presented some experimental results that seem to corroborate the idea that outstanding variations in entropy constitute salient moments in a listening experience. The results so far suggest that intra-opus information can greatly influence the perception of segmentation boundaries. It was found that most boundaries predicted by the model could be explained with Gestalt-based principles, but these principles were captured from non-annotated melodic data and reproduced in the segmentation predictions.

ACKNOWLEDGMENTS

This research was funded by EPSRC research project GR/N08049/01. Thanks to Taylan Cemgil for his advice on Markov modelling and to Marcus Pierce for helpful discussions on the listening study.

5. REFERENCES

1. Lerdahl, F., Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. M.I.T. Press, Cambridge (Mass.).
2. Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. PhD thesis, University of Edinburgh.
3. Deliège, I., Melén, M. (1997). Cue abstraction in the representation of musical form. In Deliège, I., Sloboda, J. (eds.) *Perception and Cognition of Music*. Psychology Press (pp. 387-412).
4. Meyer, L.B. (1967). *Music, The Arts, And Ideas - Patterns and Predictions in Twentieth-Century Culture*. University of Chicago Press, Chicago.
5. Huron, D. (2001). What is a musical feature? Forte's analysis of Brahms's opus 51, no. 1, revisited. *Music Theory On-line* **7**.
6. Manning, C.D., Schüttze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.
7. Conklin, D., Witten, I. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research* **24** (pp. 51-73).
8. Ponsford, D., Wiggins, G., Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research* **28**
9. Reis, Y.B. (1999). Simulating music learning: On--line, perceptually guided pattern induction of context models for multiple-horizon prediction of melodies. In *Proceedings of AISB'99 - Symposium on Musical Creativity* (pp. 58-63).
10. Bod, R. (2001). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research* **30**.
11. Saul, Lawrence, Jordan, Michael (1999). Mixed Memory Markov Models: Decomposing Complex Stochastic Processes as Mixtures of Simpler Ones. *Machine Learning*. **30** (1). (pp. 75-87).
12. Ney, Hermann and Essen, Ute and Kneser, Reinhard (1994). Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer Speech and Language*. **8** (1). (pp. 1-38).
13. Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27** (pp. 379-423,623-656).
14. Deliège, I. (1998). Wagner “alte weise”: Une approche perceptive. *Musica Scientiæ* **Special Issue** (pp. 63-90)
15. Deliège, I. (2001). Prototype effects in music listening: An empirical approach to the notion of imprint. *Music Perception* **18** (pp. 371-407).

A Probabilistic Model for Melody Segmentation

Miguel Ferrand[1]**, Peter Nelson[1], and Geraint Wiggins[2]

[1] University of Edinburgh, Scotland, UK,
{mferrand,pwn}@music.ed.ac.uk,
http://www.music.ed.ac.uk/
[2] City University, London, UK
geraint@soi.city.ac.uk,
http://www.soi.city.ac.uk/

Abstract. In this paper we propose that a probabilistic model of music listening may be used to predict segmentation boundaries in melodies,, as perceived by a listener. Existing models of music perception usually achieve a structural segmentation of a music piece based on Gestalt–based local discontinuities and on the detection of parallelism. The assimilation of regularities in music contributes to expectations created during the course of listening, and is reflected in the listener's ability (or inability) to predict what comes next. We conjecture that the expectations associated with intra-opus musical information provide strong hints for segmentation points within a piece. We describe an implementation of this model and analyse a preliminary segmentation experiment, discussing the limitations and the possible developments of this approach.

1 Introduction

When listening to music, subjects often perceive divisions in the musical discourse. The identification of several parts or segments in a piece is an important step for abstracting musical contents. Several theories have recognised music segmentation as an important part of music understanding, and have attempted to explain and formalise how listeners' intuitions account for the identification of the pieces' constituent units such as motives, phrases or sections.

Some of these theories [1–3] employ Gestalt principles to identify discontinuities or create note groupings. Although grouping principles have been found to have a reasonable explanatory power [4, 5], most theories that use Gestalt principles for segmentation often rely on higher-level rules to form larger groupings or to identify parallelisms.

Deliège and Melén [6] argued for the prototypical nature of parallelism and showed that descriptions of sections of a musical piece can be formed and retained by listeners, based on the repetition and salience of small musical patterns. These small patterns constitute indexes for larger sections and their salience is enhanced by the familiarity of the listener with the music. Familiarity with a certain musical material may result from the assimilation of a particular musical

** funded by EPSRC research project GR/N08049/01

style or repertoire, or from recent or repetitive audition and memorisation of a particular piece.

Leonard Meyer affirmed that "suspense is essentially a product of ignorance as to the future course of events" [7] outlining the relationship between acquired knowledge and expectation. Later he recognised the affinity between expectation and information theory and in particular with the notion of entropy [8]. The unpredictability of an event in a sequence of events, can alter its prominence. An unpredictable musical event is more noticeable to the listener and therefore more likely to be remembered. On the same line of thought Huron [9] argues that to establish feature salience in a musical work, one has to identify those characteristics that present intratextual or intertextual distinctiveness since "the mere presence of some element or property does not necessarily make it a good feature. A good feature must in some ways draw attention to itself".

Researchers have used the notion of entropy to model and evaluate musical composition [10, 11] or to measure musical learning [12], but to our best knowledge, none have used this concept to model music segmentation.

This paper proposes the use of a probabilistic approach to predict segmentation boundaries in melodies. We argue that musical segments are not always clearly characterised parts of a musical piece, such as those related by similarity. In the course of listening, the absence of references during a prolonged temporal interval, may lead the listener to recall that interval as a segment in the music for which no clear understanding was experienced. In other words, parts of a musical piece which, to the listener, lack an identity or a particular character, may also be identified as distinct segments.

2 Overview of the Proposed Model

2.1 General principles

We view music as a multidimensional phenomenon, where several features (e.g. pitch related or rhythm related) unfold simultaneously in time. At different moments within a piece, different musical features may be the source of discontinuities or perceived saliences. A piece of music contains more information than a listener can process in a single hearing, therefore listening implies choosing which elements to attend to, from time to time [13].

Krummansl [14] found that in a task of musical segmentation, listeners identified boundaries mainly on the basis of a combination of several different musical characteristics. In the absence of evidence to distinguish quantitatively the contribution of different musical features, we propose only to identify which ones are salient or not salient, at a particular moment. If two or more different features are found to be salient simultaneously, at a particular point, then they will add up to constitute a stronger salient moment.

In this work, and as explained previously, we associate feature salience with expectation. We will use the entropy as a measure of unpredictability associated with different musical features. Low entropy usually means high predictability

but if a particular feature (e.g. note duration) is highly predictable throughout the piece then it may well be because it is either highly invariant or because it follows a monotonous variation pattern. For example if a whole melody is layered on semi quavers, we can say that rhythm is highly predictable, but it provides no references for the segmentation of the melody. For this reason we are not interested in measuring the overall entropy of the model, but rather how entropy changes along the piece. We conjecture that transitions between high and low entropy constitute salient moments in a listening experience. Furthermore we argue that musical parameters with varying entropy along the piece are more informative than parameters with consistently high or low entropy values.

2.2 N-gram models

The implementation of the model is based on an n-gram grammar. N–gram grammars are n^{th} order Markov models that assume that the probability of occurrence of a symbol depends on the prior occurrence of $n-1$ other symbols. N–gram models are typically constructed from statistics obtained from a large corpus of data (usually referred to as the training corpus) using the co-occurrences of symbols to determine the probabilities of sequences of symbols.

Hence, the probability $P(s)$ of a sequence $s = w_1...w_l$ of length l is given by,

$$P(s) = \prod_{i=1}^{l} P(w_i | w_{i-n+1}^{i-1}) \tag{1}$$

where w_i^j denotes the sequence $w_i...w_j$ and n is the order of the model [3].

Independently of the size of the training corpus, it is unlikely that all possible symbol sequences will occur. Data sparseness becomes a problem if, when computing probabilities using Equation 1 some of the terms in the product have zero probability. Also, if the training corpus is small, and the order of the model is significantly high, longuer sequences will have relatively lower counts, resulting in less accurate probabilities.

Several methods, usually referred to as *smoothing* methods, have been described in the literature [15, 16] to overcome the data sparseness problem, and estimate probabilities. In this work we are focusing only on intra-opus information meaning that the amount of data to be analysed is substantially lower than if we were using a larger corpus of pieces, so a linear interpolation smoothing method [17] was employed. Using linear interpolation the probabilities of a sequence of length l can be estimated by a weighted sum of n–gram probabilities from models of order $n <= l$. For instance, the probability of a tri-gram is determined by the weighted sum of corresponding uni-gram, bi-gram and tri-gram probabilities,

$$P(w_k | w_{k-2}, w_{k-1}) = \lambda_1 P(w_k) + \lambda_2 P(w_k | w_{k-1}) + \lambda_3 P(w_k | w_{k-2}, w_{k-1}) \tag{2}$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1 < \lambda_2 < \lambda_3$ as it is assumed that longer contexts, being more specific, should have a higher weight.

[3] when $n > i$ padding symbols have to be introduced to provide the necessary contexts

2.3 Entropy

The fundamentals of Information Theory (IT) were first introduced in [18], and set up quantitative ways of measuring the information contained in a message being transmitted, received, or stored. One of the ways of measuring the quantity of information of a particular message is to determine its unpredictability or entropy. For a given N-gram model M, entropy associated with a given context c can be determined by,

$$H_c(M) = -\sum_{\forall e:(c,e)\in M} P(e|c)\log_2 P(e|c) \tag{3}$$

where e denotes all possible successor symbols of the context c. Contexts are sequences of size $N-1$ where N is the order of the model M. Conditional probabilities are calculated using Equation 2.

Since we are interested is observing the changes in entropy along a sequence of symbols, the occurrence of every new symbol in the sequence provides a new context for which the values of $H_c(M)$ can be calculated.

3 A case study

Seeking to compare the proposed model with segmentation data provided by real listeners we used some data described in a segmentation experiment carried out by Deliège [19]. In these experiments subjects listened to a melody (the solo for English Horn, from Wagner's opera Tristan and Isolde) and had to identify segmentation points in real–time. Both musically trained and untrained subjects took part in the experiments. A familiarisation audition of the piece preceded the auditions during which subjects were asked to identify segmentation boundaries. The experiments revealed a set of 8 main segment boundaries (identified by most subjects) and an additional set of 13 weaker boundaries. For the present study only the stronger boundaries were used for comparison. For full details of the experimental procedure the reader is referred to [19].

The melody information was translated into an event–based representation. All events are numbered sequentially and gather information about pitch (Midi note code), duration and onset time. From these basic event attributes, four other features were extracted and associated with each event:

- Pitch step (PS): expresses the interval distance to following event in semi–tones.
- Pitch contour (PC): expresses the sign of the pitch step; takes value -1,+1 or 0 if PS is also 0.
- Duration ratio (DR): expresses the ratio between the durations of the present and the following event.
- Duration contour (DC): expresses duration ratio changes; takes values -1 if $DR > 1$, 1 if $DR < 1$ or 0 if $DR = 1$.

Table 1. The first 14 events of the melody and features extracted: pitch step (PS); pitch change (PC); duration contour (DC) and duration ratio (DR)

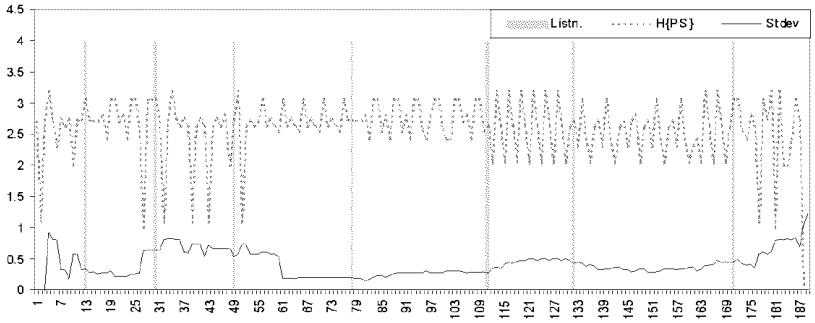
No	Midi	Dur	OnSet	PS	PC	DC	DR
1	53	2.000	0.000	7	1	1	1.250
2	60	2.500	2.000	3	1	-1	0.200
3	63	0.500	4.500	-2	-1	1	3.000
4	61	1.500	5.000	-5	-1	-1	0.333
5	56	0.500	6.000	7	1	1	3.000
6	63	1.500	6.500	-8	-1	-1	0.333
7	55	0.500	8.000	5	1	0	1.000
8	60	0.500	8.500	-7	-1	0	1.000
9	53	0.500	9.000	5	1	1	3.000
10	58	1.500	9.500	2	1	-1	0.333
11	60	0.500	10.500	-2	-1	1	4.000
12	58	2.000	11.000	-2	-1	0	1.000
13	56	2.000	13.000	-1	-1	-1	0.167
14	55	0.333	15.000	-2	-1	0	1.000
15	53	0.333	15.333	-2	-1	0	1.000
..	..	...	...	.	.	.	...

Table 1 shows an extract of the encoded melody of the Horn solo, with the additional four attributes.

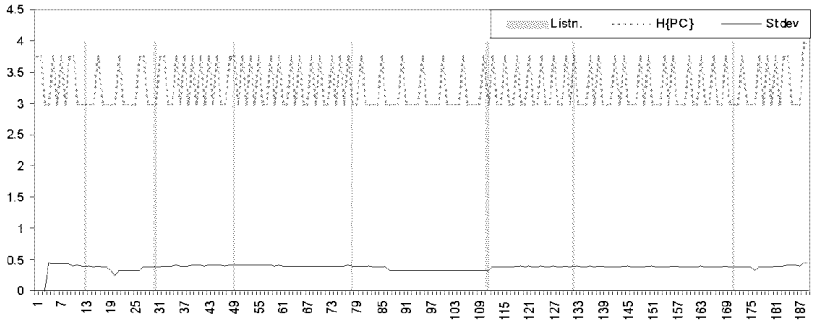
The values obtained for the attributes PS, PC, DC and DR, constitute four sequences of symbols from which sequence probabilities can be generated. A tri–gram, bi–gram and uni–gram model was generated for each one of the four sequences.

The entropy values were obtained from each one of the models, and for all events in the melody. In Figure 1 we show the entropy profiles obtained from a tri–gram model. The vertical grey lines overlapped in the graph indicate the locations of the stronger boundaries indicated by the listeners in Deliège's experiment [19]. The standard deviation (*stdev*) of the entropy is also depicted in the lower part of each graph. Standard deviation gives a good measure of the spread of the entropy values along the graph and since we are interested in measuring the changes in entropy along the piece, the *stdev* is calculated with a sliding window. In this experiment we used a fixed size window of 10 events although we suggest that the size of window could be determined in terms of time and not in number of events. This would seem perceptually more realistic and the sliding window, with a fixed duration, could be seen as a short–term memory time frame within which changes can be perceived by the listener. The number of events that would fit in this window would depend on the tempo assigned to the piece. Further research is necessary to corroborate this idea.

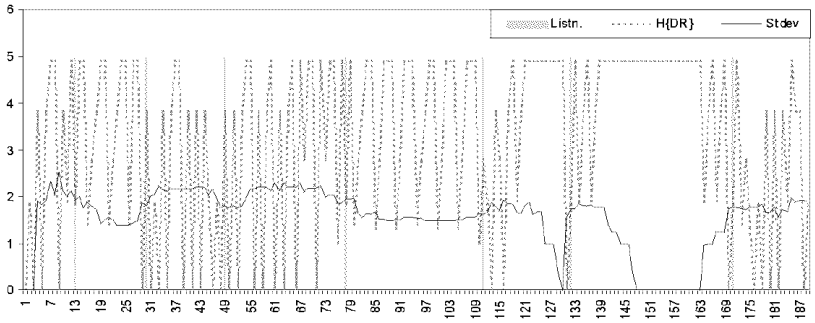
In a first observation of the graphs of Figure 1 it seems clear that duration based features register a much higher entropy variance along the melody than pitch based features. Following our conjecture, time based features are then



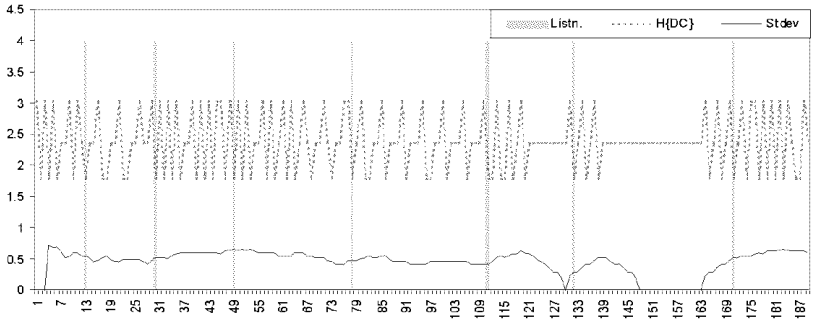
a) Pitch step entropy



b) Pitch contour entropy



c) Duration ratio entropy



d) Duration contour entropy

Fig. 1. Entropy(dotted line) and entropy moving standard deviation (solid line) for PS, PC, DC and DR *vs.* event number. Listeners' segmentation boundaries are indicated by grey vertical lines

likely to convey more information to the listener regarding segmentation. In fact it can be observed that the DR entropy graph exhibits accentuated variations, many of which occur in the vicinity of the boundaries indicated by listeners. This is confirmed by Deliège in the analysis of her experimental results, where it is reported that the listeners' decision were dominated by Gestalt principles of proximity which are more evident in the rhythmic content of this melody. The graphs of PS and PC, although with overall low *stdev* display different entropy variation patterns within many of the segments identified by the listeners in Deliège's experiment.

At present we do not yet have a method to automatically interpret and extract these boundaries from the entropy profiles obtained from the melodies. This part of the model will be contemplated in further developments.

4 Discussion

The use of n–gram models is often criticised for the underlying assumption that a state depends only on the previous states. This assumption seems to be over–simplistic if we are analysing musical sequences, however it is known that human memory limitations impose a limit on the ability to establish large–span temporal relations. This has been acknowledged in [1, 20] where it has been suggested that listeners perceive a musical surface by focusing on successive zones, along the musical piece.

As mentioned before, parallelism has a strong influence in the establishment of boundaries in a melody so the perception of similarities cannot be ignored if we want to model segmentation. A probabilistic model, like the one presented here, can capture some parallelisms but it is limited by the order of the model, which determines the length of the patterns that can be stored and recalled as similar. Because only identical sequences are recorded as repetitions, this probabilistic model can, in principle, only deal with exact similarity. However, it can be argued that because several features were separated into more general descriptors (e.g. duration ratio or pitch contour as opposed to absolute duration or pitch values) we can accommodate some form of approximate similarity. For example two sequences can have the same relative durations, although different absolute note durations. Although large patterns cannot be stored as a whole by a low order model, we argue that parallelism may still be established based on smaller parts of these patterns. Empirical studies [21] have shown that primacy effects were present in the recognition of similar patterns, and that patterns are often remembered by the repetition of smaller patterns, which are often their initial sections. This evidence supports the conjecture that a short–term memory model can capture some structural information based on parallelism, provided there is some regularity in the acquired musical data.

In this work, only intra-opus information was used, meaning that the model only capture regularities within a given piece. The results shown were obtained with a model of order 3. As expected, and due to the fairly small set of input data, increasing the size of the sequences stored by the model decreases the

overall pattern count, and therefore compromising probability estimation. The use of interpolation and the unfolding of the melodic information in several features provided additional redundancy but not enough to accommodate very long context models. Additional experiments are necessary to find out how the order of the model influences the granularity of the entropy profiles.

The use of inter–opus in conjunction with intra-opus information was suggested in [10], where two separate models are combined to make predictions, although the way these two context models were actually combined has not been described in detail. Intuitively, it seems that regularities particular to a musical piece could override the intuitions resulting from long term established rules. The long–term model provides the 'norms' (obtained from a large corpus of pieces) and the short–term model, obtained from the piece being heard, provides the listener with confirmations or deviations from the rules.

4.1 Related work

An important contribution of the present approach is that it attempts to predict segmentation boundaries from non–annotated musical data. Most other related approaches include style–dependent knowledge or use pre-annotated training data.

Conklin and Witten's [10] multiple viewpoint system for generation of Bach chorales uses a training corpus which includes score based information such as time signatures, fermatas, location of bar lines, *etc.* Ponsford et al. [11] use a probabilistic approach based on N-grams to capture and generate harmonic sequences, but also assume from the start the use of a score–based music representation.

Bod [22] proposes the use of a Markov Grammar to learn and predict phrase boundaries in folk songs. Learning is based on a training set of pre-annotated pieces obtained from the Essen folk song database. The phrase boundaries indicated in the Essen folk songs have not been validated with listeners, and the author acknowledges that the correction of the annotations should preferably be established by an independent psychological experiment with more than one subject. This raises the question whether the model is really predicting listening behaviour or just predicting boundaries according to particular analytical criteria, reflected in the annotations of the pieces in the database? To answer this question it would be necessary to test a sample of songs from the Essen database with listeners, and find out how the phrase boundaries indicated by listeners would differ from the annotated ones. In short and simple pieces, where parallelism is more obvious, it is likely that the structural segments perceived by listeners correspond to the sections obtained by simple analysis of the pieces. However in longer pieces, and when parallelism is more difficult to establish, either by the temporal separation between motives or by the subtlety of the similarities, models should be compared with results obtained by listeners.

Reis' [12] research aimed to determine to what degree a system without any *a priori* stylistic information, is able to gain proficiency in a given musical style, as measured by its ability to predict the music. The author extends the approach

based on context models [10] to simulate the on–line process of music learning and capture the stylistic information present in a musical surface. He argues for a more cognitively pertinent way of inducing the contexts, using Gestalt–based perceptual cues (e.g. changes of direction, or large jumps in either pitch or time domain) to restrict the number of sequences that are extracted and stored from the training set.

There are advantages in including contextual information in an analytical process. For example, Bod [22] has shown how to improve the performance of his model by limiting the maximum number of phrase boundaries the parser can identify within a song (this maximum may be obtained directly from the Essen database). The drawback of this sort of approach is that the model becomes too biased towards a particular repertoire, so it is likely that the predictive power may drop when parsing pieces from other repertoires. In fact, in the particular case of the Essen folk song database, it is likely that the number of segments may vary significantly according to the origin or type of the songs.

5 Conclusions

This study suggests that some structural information about a melody can be associated and induced by changes in expectation. It was found that distinct changes in entropy associated with different musical features were coincident with melody segment boundaries indicated by listeners. It was also shown that the statistical properties of the entropy profiles may be used to indicate which parts of a melody, or more generally which features of a melody are more informative and therefore more likely to contribute to the perception of segmentation boundaries.

A central motivation of this work is to develop a model that can predict segmentation boundaries by learning from non–annotated data. Preliminary results reveal that the model has a significant predictive power, concerning the location of segmentation boundaries and thus encourages further developments and experimental research.

References

1. Lerdahl, F., Jackendoff, R.: A Generative Theory of Tonal Music. M.I.T. Press, Cambridge (Mass.) (1983)
2. Narmour, E.: The Analysis and Cognition of Basic Melodic Structures: The Implication–Realisation Model. University of Chicago Press, Chicago (1990)
3. Cambouropoulos, E.: Towards a General Computational Theory of Musical Structure. PhD thesis, University of Edinburgh (1998)
4. Deliège, I.: Grouping conditions in listening to music: an approach to Lerdahl and Jackendoff's grouping preference rules. Music Perception **4** (1987) 325–360
5. Bigand, E., Lerdhal, F., Pineau, M.: Deux approches expérimentales des quatre composants de la théorie generative de la musique tonale. In Deliége, I., ed.: Proceedings of the 3rd International Conference on Music Perception and Cognition, Liège, Belgium, European Society for the Cognitive Sciences of Music (1994) 259–260

6. Deliège, I., Melén, M.: Cue abstraction in the representation of musical form. In Deliège, I., Sloboda, J., eds.: Perception and Cognition of Music. Psychology Press (1997) 387–412
7. Meyer, L.B.: Emotion and Meaning in Music. University of Chicago Press (1956)
8. Meyer, L.B.: Music, The Arts, And Ideas – Patterns and Predictions in Twentieth–Century Culture. University of Chicago Press, Chicago (1967)
9. Huron, D.: What is a musical feature? Forte's analysis of Brahms's opus 51, no. 1, revisited. Music Theory On-line **7** (2001)
10. Conklin, D., Witten, I.: Multiple viewpoint systems for music prediction. Journal of New Music Research **24** (1995) 51–73
11. Ponsford, D., Wiggins, G., Mellish, C.: Statistical learning of harmonic movement. Journal of New Music Research **28** (1999)
12. Reis, Y.B.: Simulating music learning: On–line, perceptually guided pattern induction of context models for multiple-horizon prediction of melodies. In: Proceedings of AISB'99 - Symposium on Musical Creativity. (1999) 58–63
13. Aiello, R.: Can listening to music be experimentally studied? In Aiello, R., Sloboda, J., eds.: Musical Perceptions. Oxford University Press (1994) 273–282
14. Krumhansl, C.L.: Cognitive Foundations of Musical Pitch. Number 17 in Oxford Psychology Series. Oxford University Press, Department of Psychology, Cornel University (1990)
15. Manning, C.D., Schüttze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, Mass. (1999)
16. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modelling. In: Proceedings of the 34th Annual Meeting of the ACL. (1996)
17. Jelinek, F., Mercer, R.: Interpolation estimation of Markov source parameters for sparse data. Pattern Recognition in Practice (1980) 381–397
18. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379–423,623–656
19. Deliège, I.: Wagner "alte weise": Une approche perceptive. Musica Scientiæ**Special Issue** (1998) 63–90
20. Bigand, E.: Contributions of music to research on human auditory cognition. In McAdams, S., Bigand, E., eds.: Thinking in Sound: The Cognitive Psychology of Human Audition. Oxford University Press (1993) 231–277
21. Deliège, I.: Prototype effects in music listening: An empirical approach to the notion of imprint. Music Perception **18** (2001) 371–407
22. Bod, R.: Memory–based models of melodic analysis: Challenging the gestalt principles. Journal of New Music Research **30** (2001)